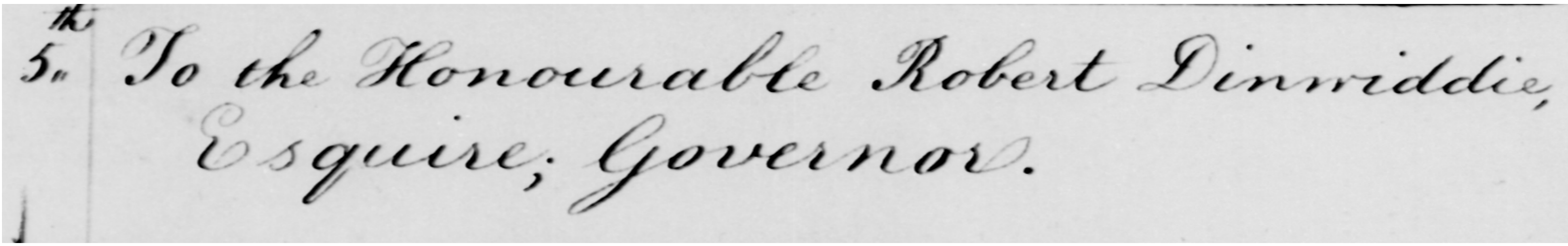
Part-Structured Inkball Models for One-Shot Handwritten Word Spotting

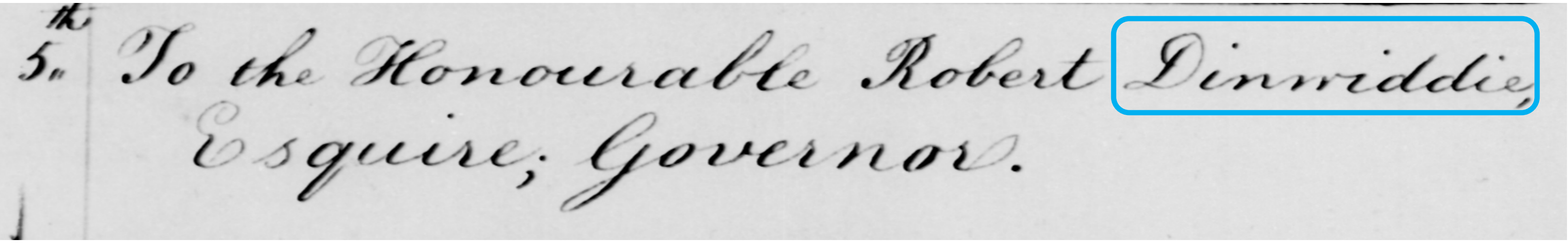
Nicholas R. Howe

Smith College

Word Spotting (by Example)

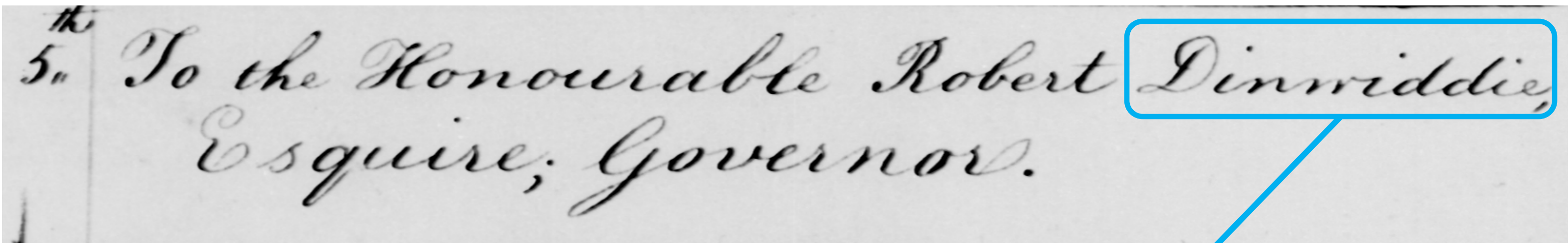

5th To the Honourable Robert Dinwiddie,
Esquire; Governor.

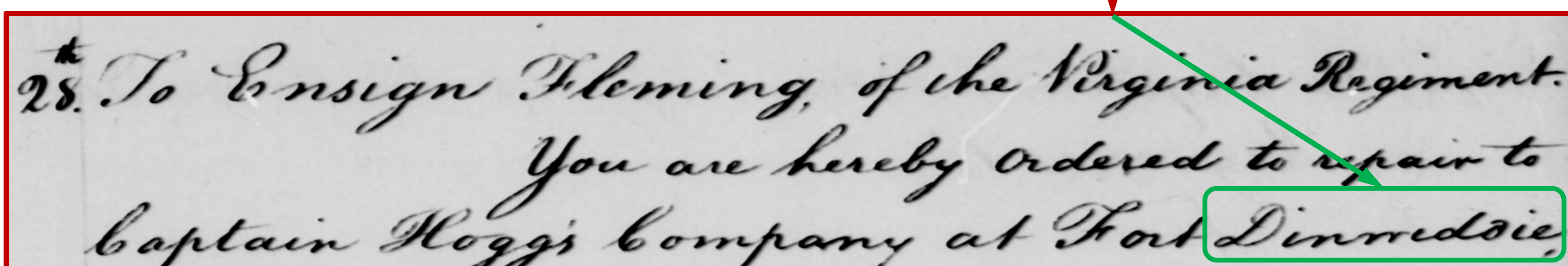
Word Spotting (by Example)


5th To the Honourable Robert Dinwiddie
Esquire, Governor.

Word Spotting (by Example)

Word Spotting (by Example)

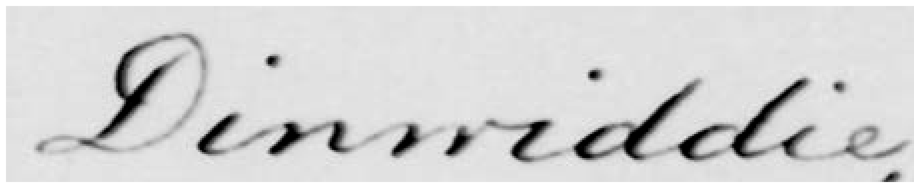

5th To the Honourable Robert Dinwiddie
Esquire; Governor.


28th To Ensign Fleming of the Virginia Regiment.
You are hereby Ordered to repair to
Captain Hogg's Company at Fort Dinwiddie

… etc.

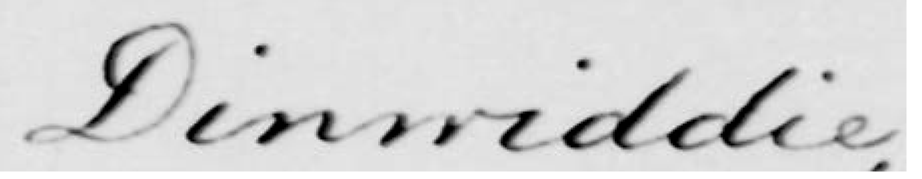
One-Shot Learning

Single example is all you get (usually)

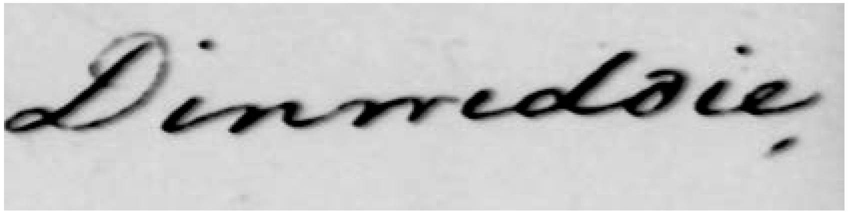


One-Shot Learning

Single example is all you get (usually)

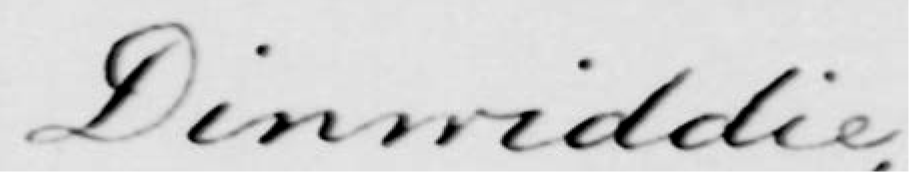


Handwriting varies – must generalize to match

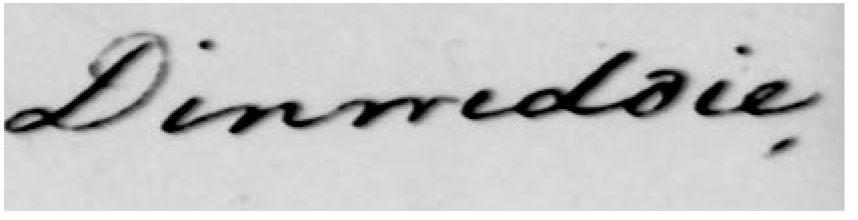


One-Shot Learning

Single example is all you get (usually)



Handwriting varies – must generalize to match



Flexibility is essential – no planar transformations

Part-Structured Models

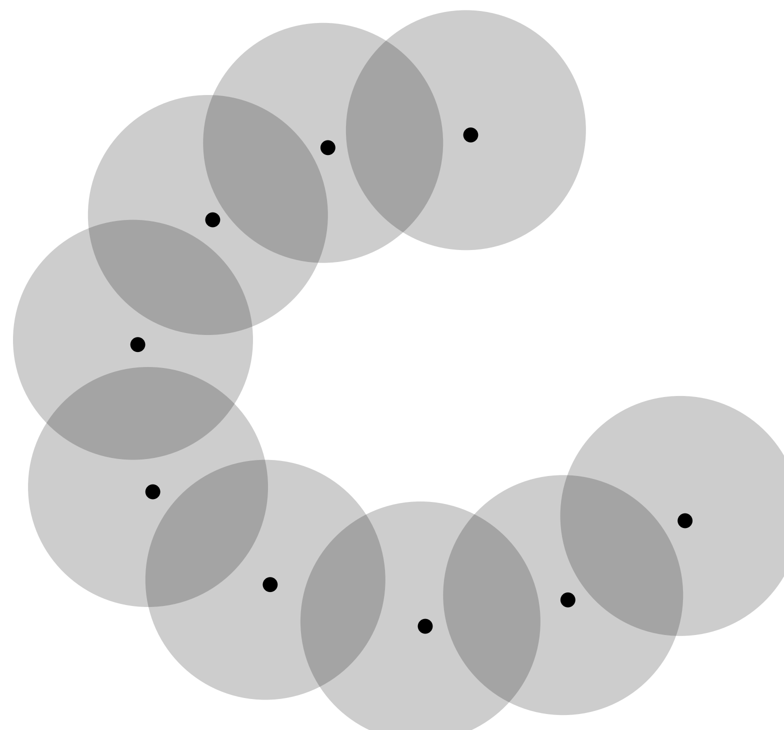
- Used for photographic object recognition
- Detected parts arranged in approximate spatial configuration

Part-Structured Models

- Used for photographic object recognition
- Detected parts arranged in approximate spatial configuration
- Successful fit identifies required parts near expected position

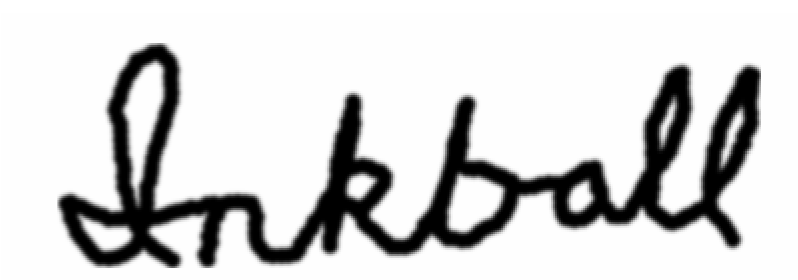
Inkball Models

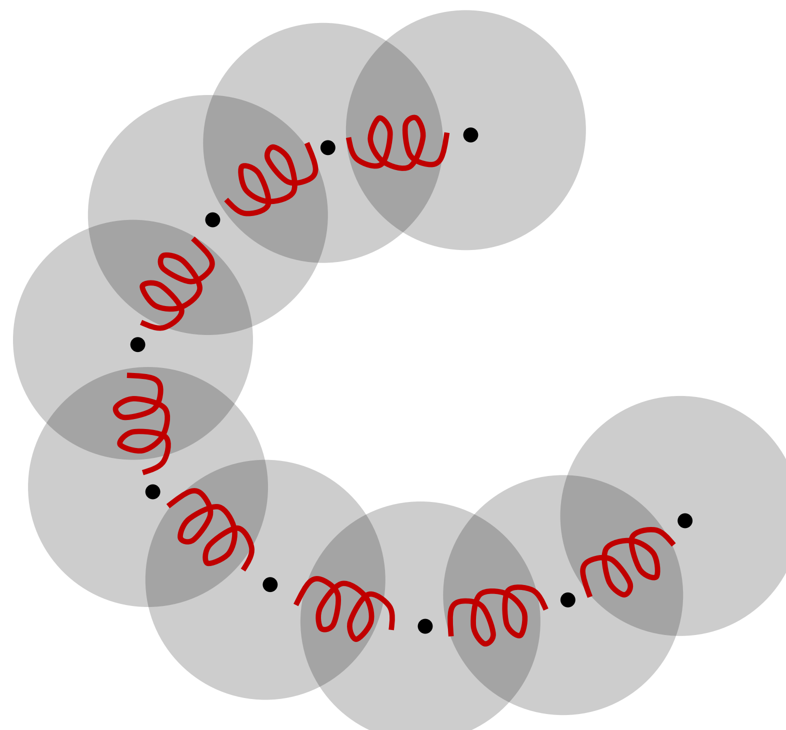
- Model = Closely spaced inkballs forming curve
- Part = Ball of ink
- Tree structure



Inkball Models

- Model = Closely spaced inkballs forming curve
- Part = Ball of ink
- Tree structure
- Connections are flexible links





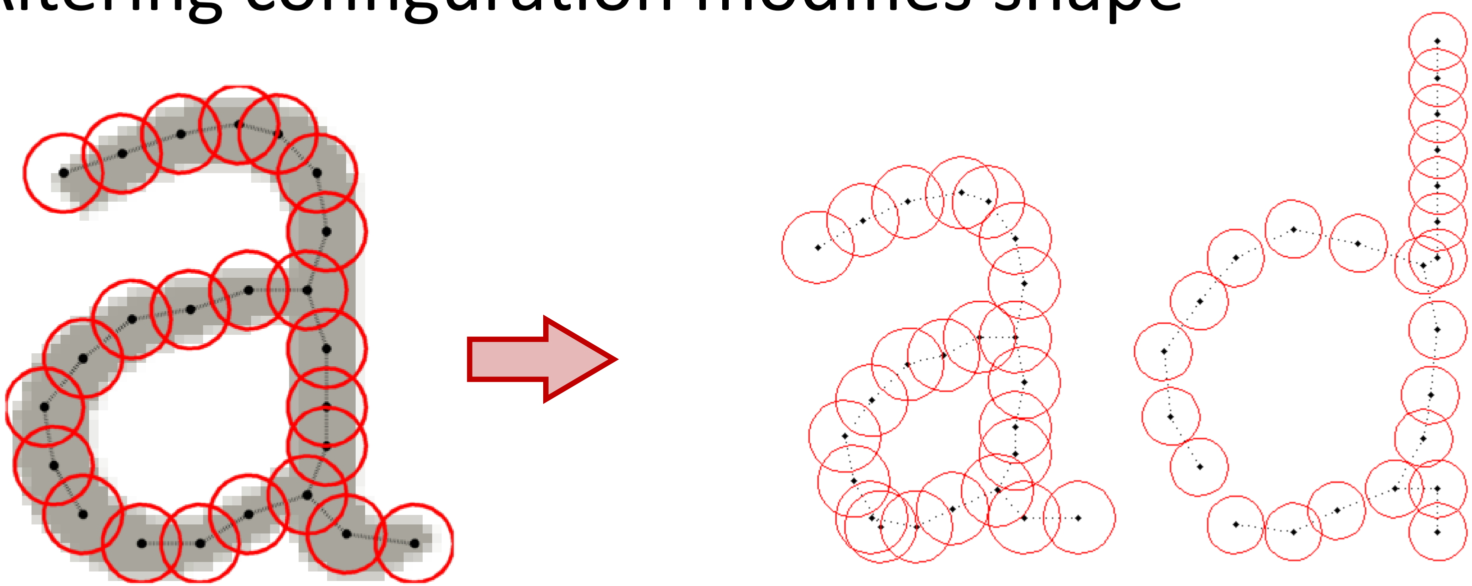
Part-Structured Inkball Models for One-Shot Handwritten Word Spotting

So, now you know.

…but how do we use these models for word spotting?

Configurations

- Configuration = 2D position for each ball
- Rest/default configuration derived from example
- Altering configuration modifies shape



Rest Configuration

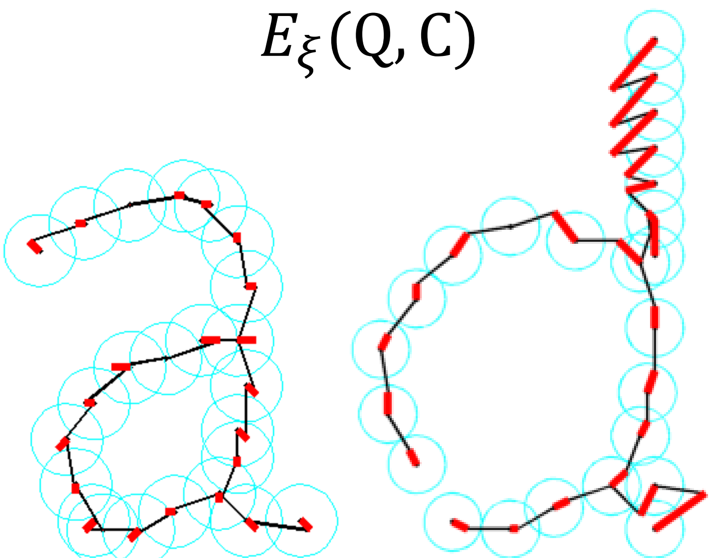
Alternate Configurations

Configuration Energy

- Match of model to image has two terms:

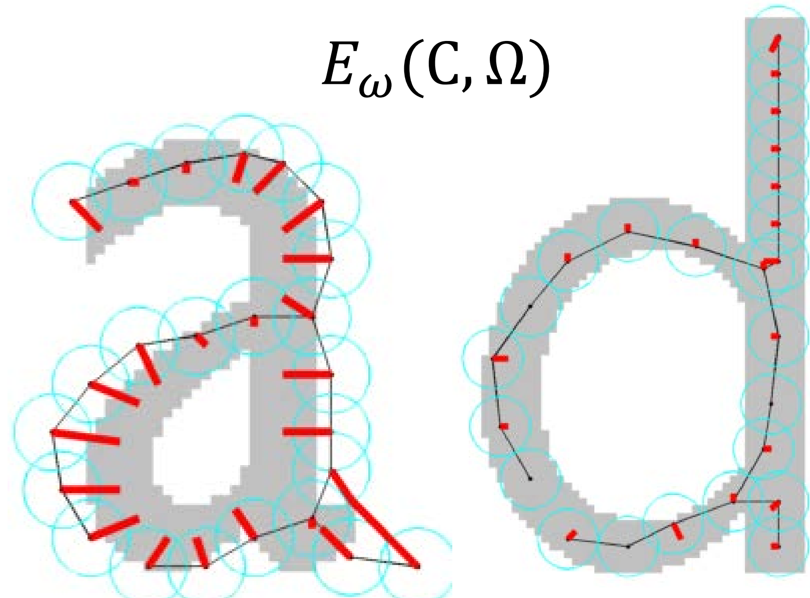
Internal deformation:
how far from default?

$E_{\xi}(\mathrm{Q}, \mathrm{C})$



Observational deformation:
how far from ink skeleton?

$E_{\omega}(\mathrm{C}, \Omega)$



Configuration Energy

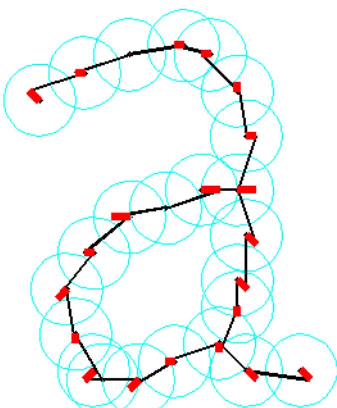
- Match of model to image has two terms:

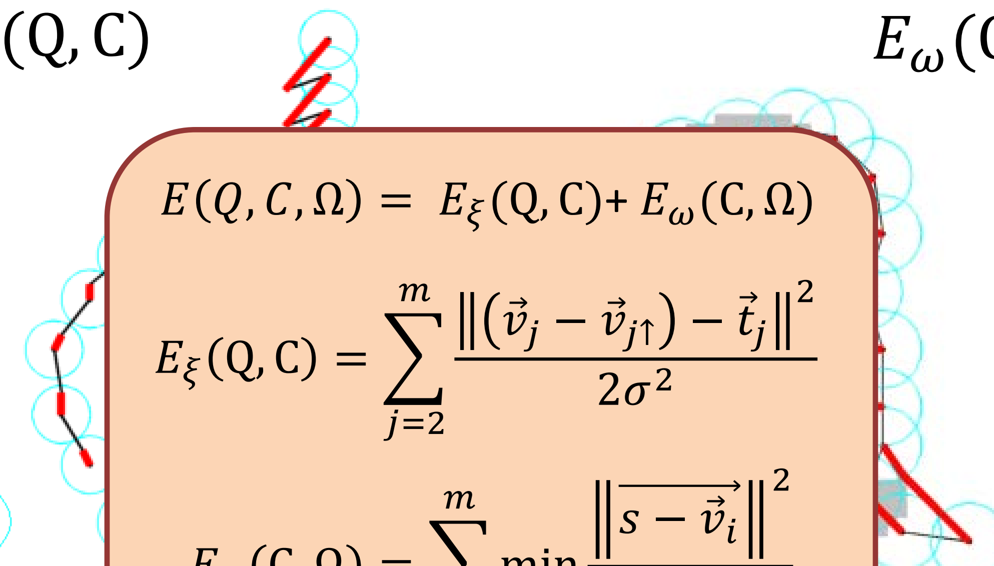
Internal deformation: how far from default?

$E_{\xi}(\mathrm{Q},\mathrm{C})$

Observational deformation: how far from ink skeleton?

$E_{\omega}(\mathrm{C},\Omega)$

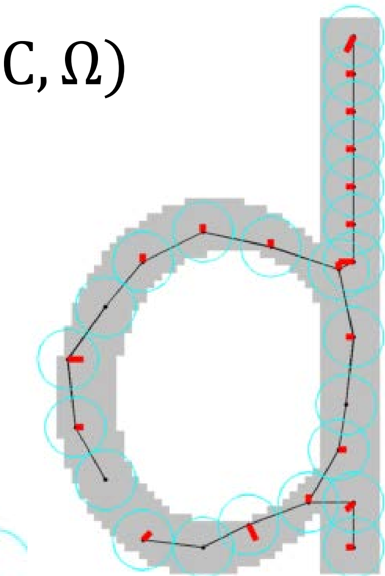




$$E(Q,C,\Omega) = E_{\xi}(\mathrm{Q},\mathrm{C}) + E_{\omega}(\mathrm{C},\Omega)$$

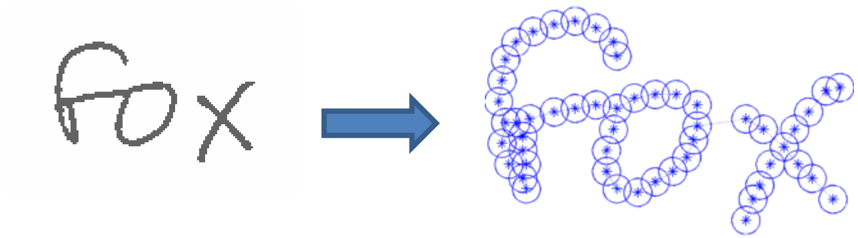
$$E_{\xi}(\mathrm{Q},\mathrm{C}) = \sum_{j=2}^{m} \frac{\left\| \left(\vec{v}_j - \vec{v}_{j\uparrow}\right) - \vec{t}_j \right\|^2}{2\sigma^2}$$

$$E_{\omega}(\mathrm{C},\Omega) = \sum_{i=1}^{m} \min_{\vec{s}\in S} \frac{\left\| \overrightarrow{s - \vec{v}_i} \right\|^2}{2\acute{\sigma}_i{}^2}$$

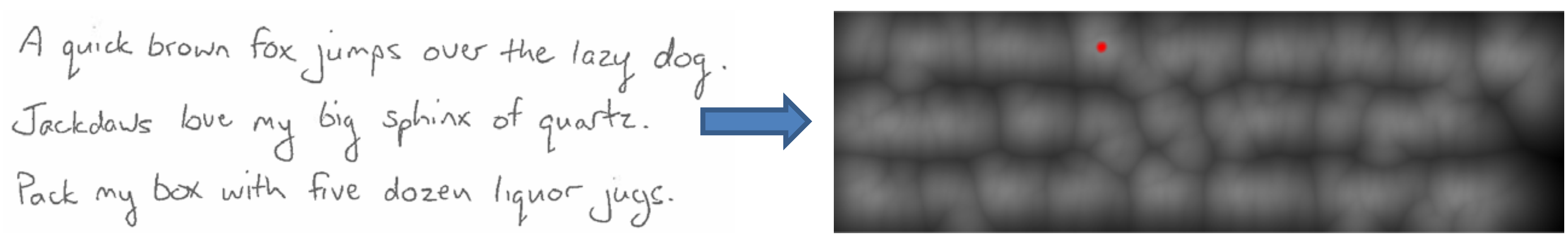


One-Shot Word Spotting

1. Infer inkball model from word sample

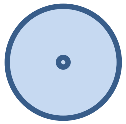


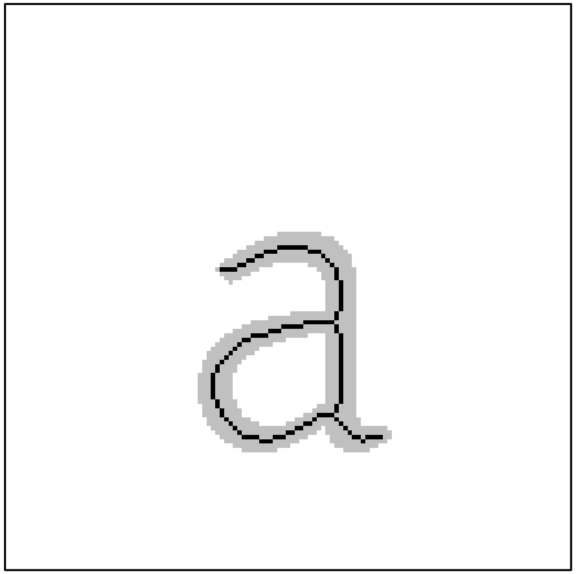
2. Efficiently identify model configurations with low energy in target document



3. Confirm candidates via reverse match

Efficient Energy Minimization

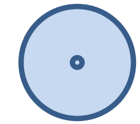
- Consider simplest case: single-node model 
  - Observation deformation is only term in play
  - Compute the energy for all possible configurations
    Distance to closest ink is just a distance transform



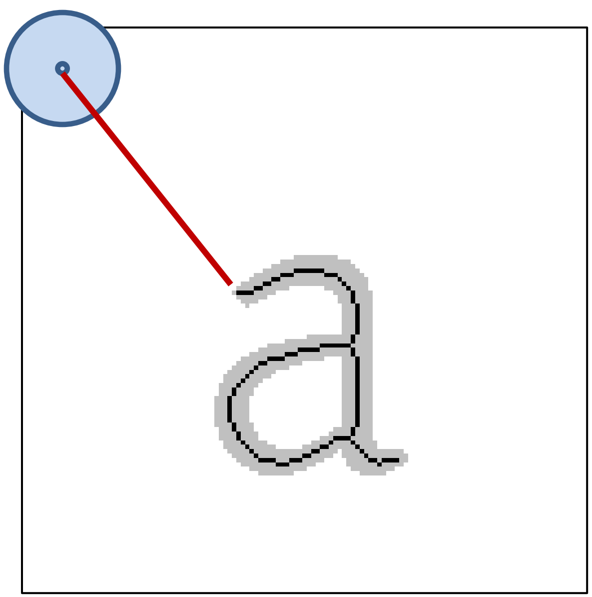
Target image

Efficient Energy Minimization

- Consider simplest case: single-node model

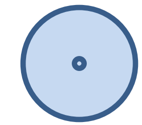


  - Observation deformation is only term in play
  - Compute the energy for all possible configurations
    Distance to closest ink is just a distance transform

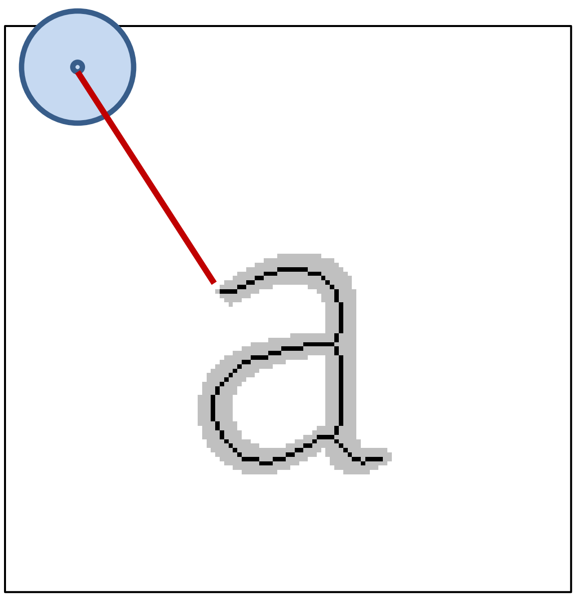


Target image

Efficient Energy Minimization

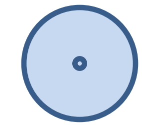
- Consider simplest case: single-node model 
  - Observation deformation is only term in play
  - Compute the energy for all possible configurations

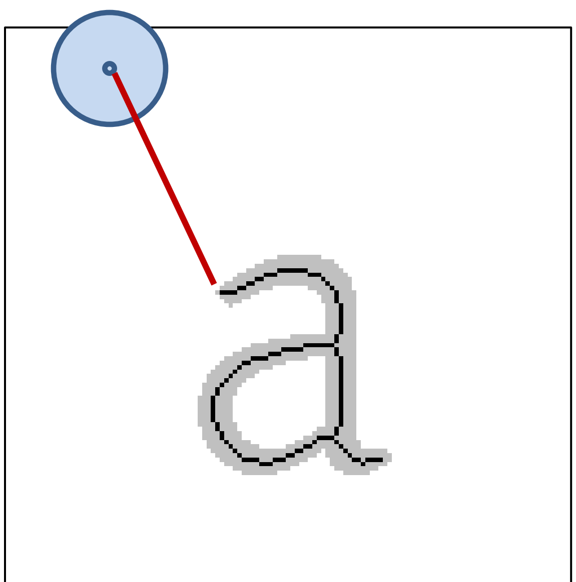
    Distance to closest ink is just a distance transform



Target image

Efficient Energy Minimization

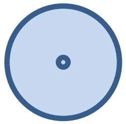
- Consider simplest case: single-node model 
  - Observation deformation is only term in play
  - Compute the energy for all possible configurations
    Distance to closest ink is just a distance transform



Target image

Efficient Energy Minimization

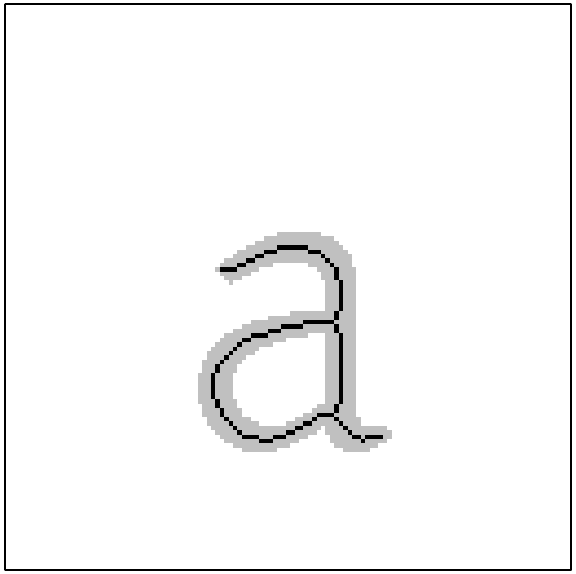
- Consider simplest case: single-node model



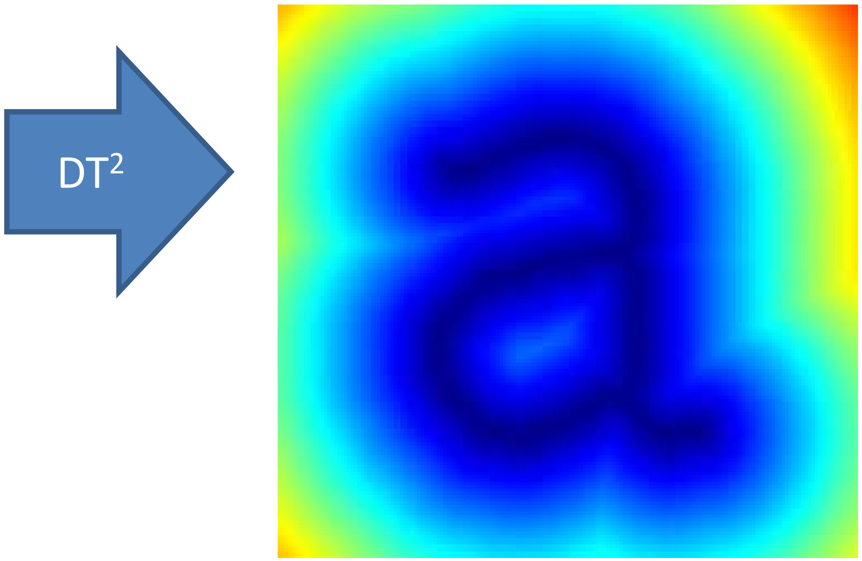
  – Observation deformation is only term in play

  – Compute the energy for all possible configurations

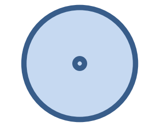
  Distance to closest ink is just a distance transform



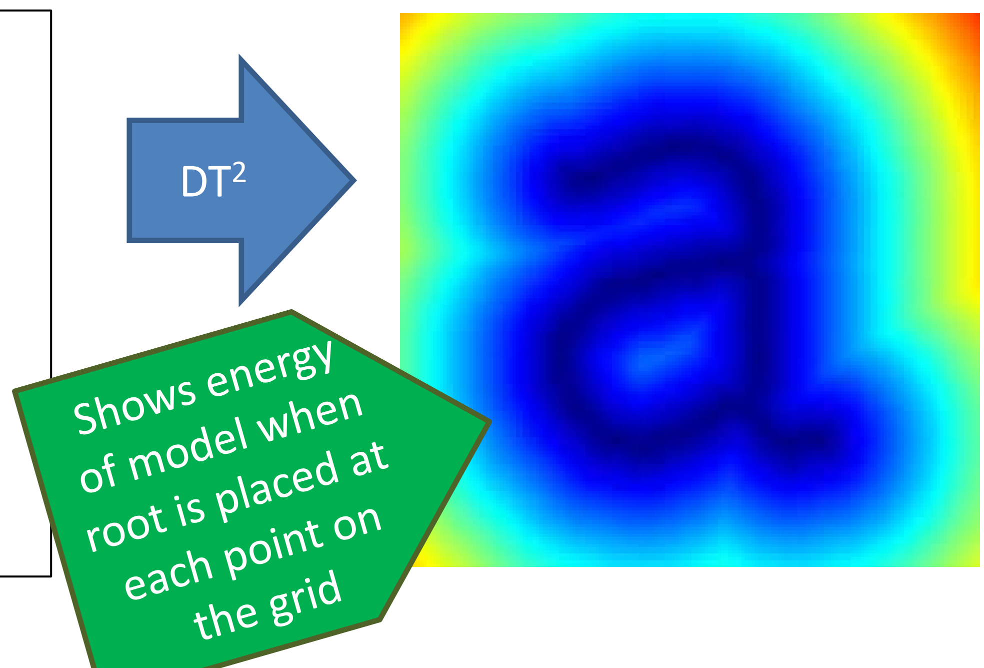
Target image



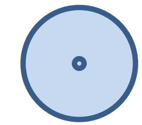
Efficient Energy Minimization

- Consider simplest case: single-node model 
  - Observation deformation is only term in play
  - Compute the energy for all possible configurations
    Distance to closest ink is just a distance transform

Target image



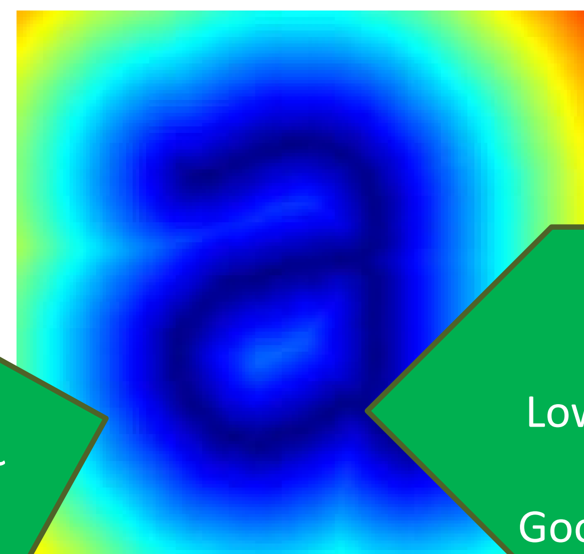
Efficient Energy Minimization

- Consider simplest case: single-node model 
  - Observation deformation is only term in play
  - Compute the energy for all possible configurations

    Distance to closest ink is just a distance transform

Target image

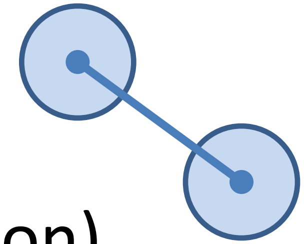
DT²



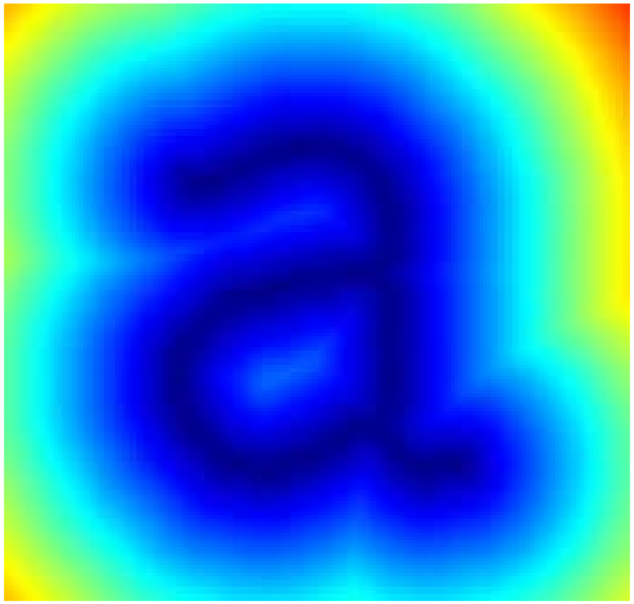
Shows energy of model when root is placed at each point on the grid

Blue == Low energy == Good match

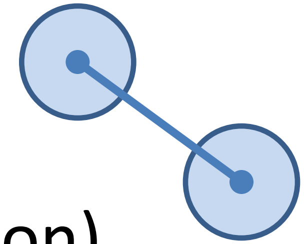
Efficient Energy Minimization



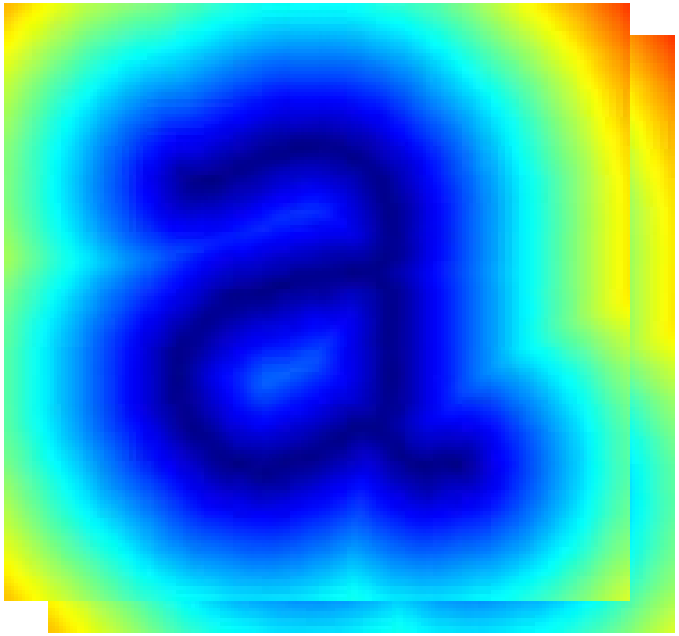
- Slightly harder case: **barbell model**
  - Still observation terms only (fixed separation)
  - Energy is sum of offset distance transforms:



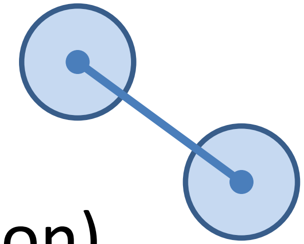
Efficient Energy Minimization



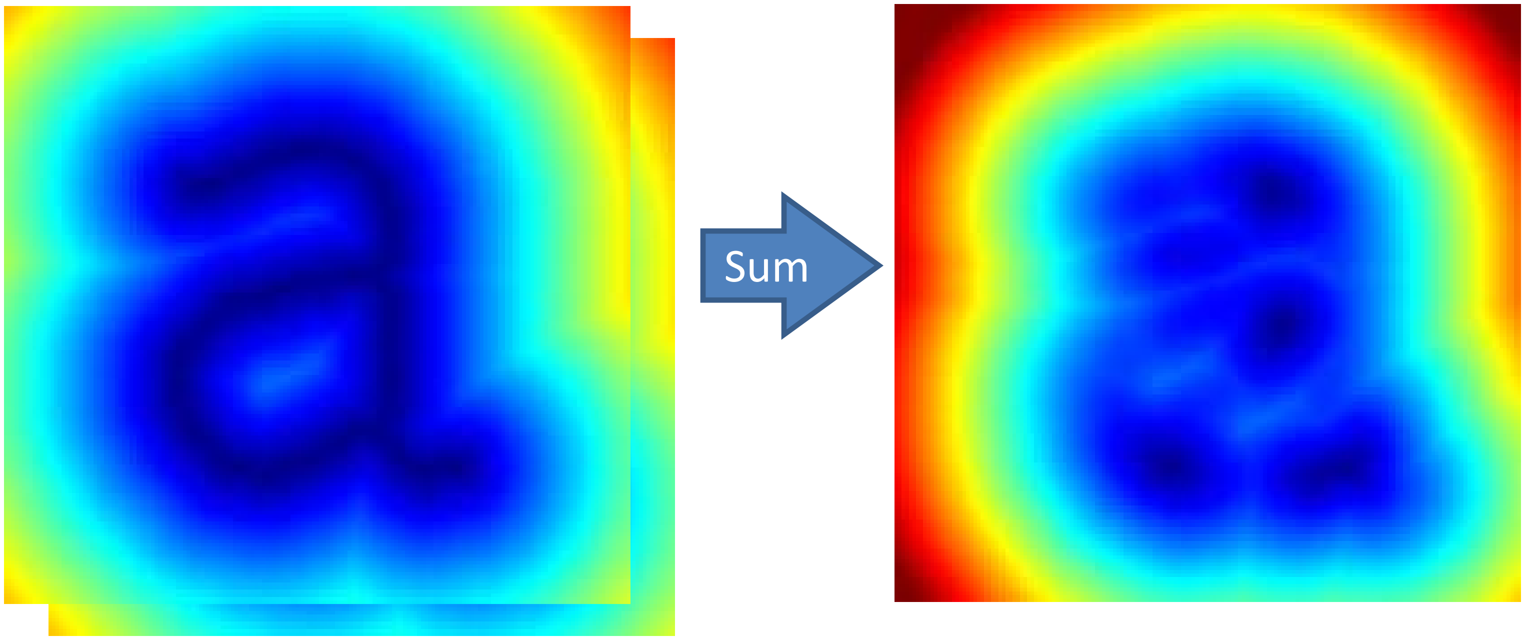
- Slightly harder case: **barbell model**
  - Still observation terms only (fixed separation)
  - Energy is sum of offset distance transforms:



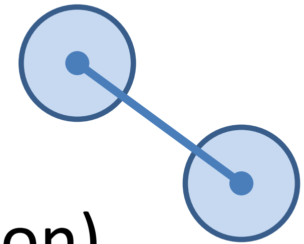
Efficient Energy Minimization



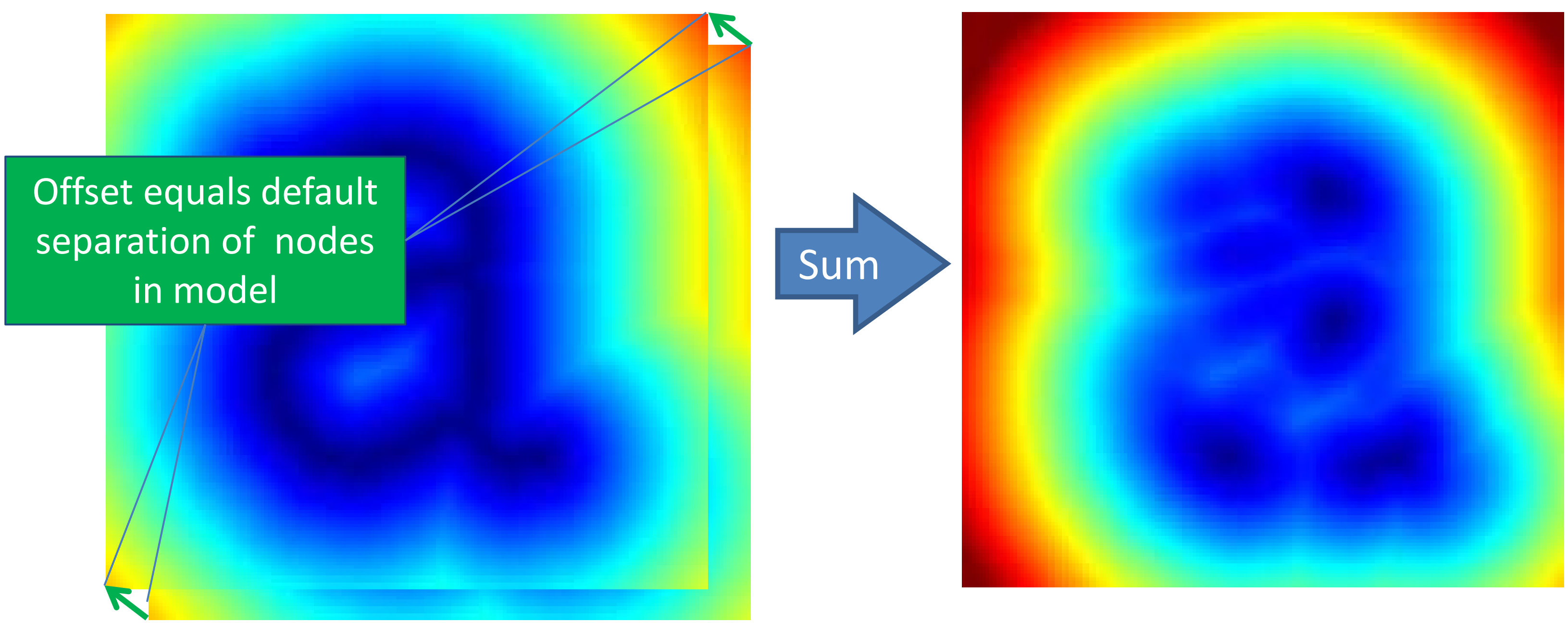
- Slightly harder case: **barbell model**
  - Still observation terms only (fixed separation)
  - Energy is sum of offset distance transforms:



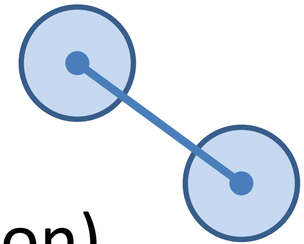
Efficient Energy Minimization



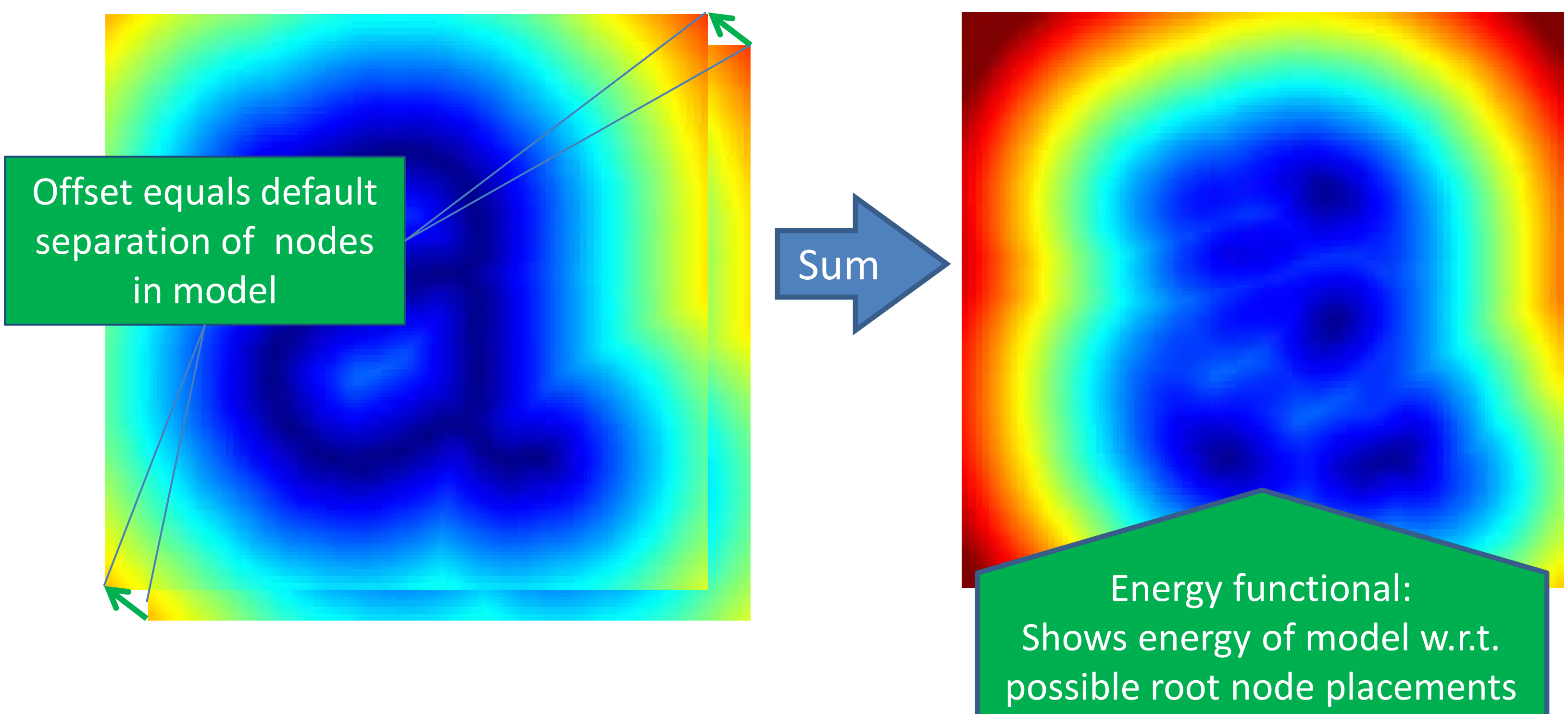
- Slightly harder case: **barbell model**
  - Still observation terms only (fixed separation)
  - Energy is sum of offset distance transforms:



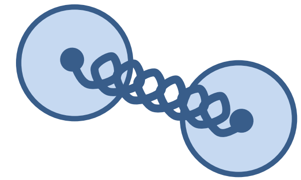
Efficient Energy Minimization



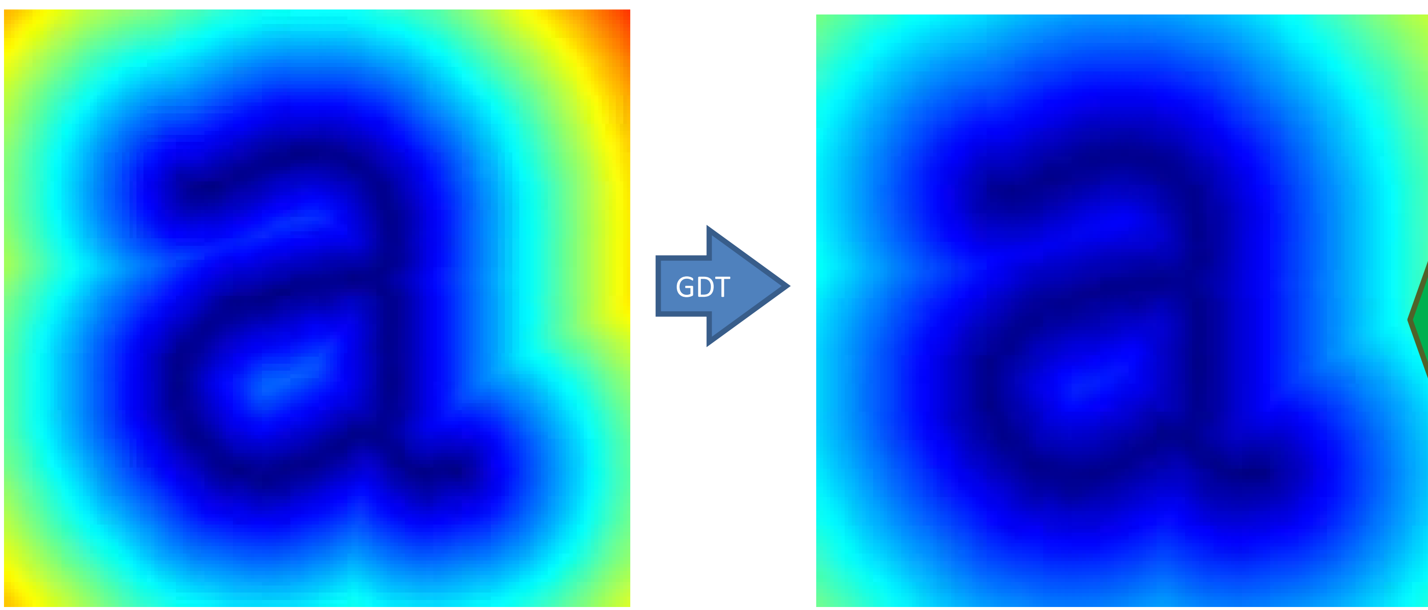
- Slightly harder case: **barbell model**
  - Still observation terms only (fixed separation)
  - Energy is sum of offset distance transforms:



Efficient Energy Minimization



- More complication: springy barbell
  - Internal deformation term enters picture
  - Use *generalized distance transform* on offset energy

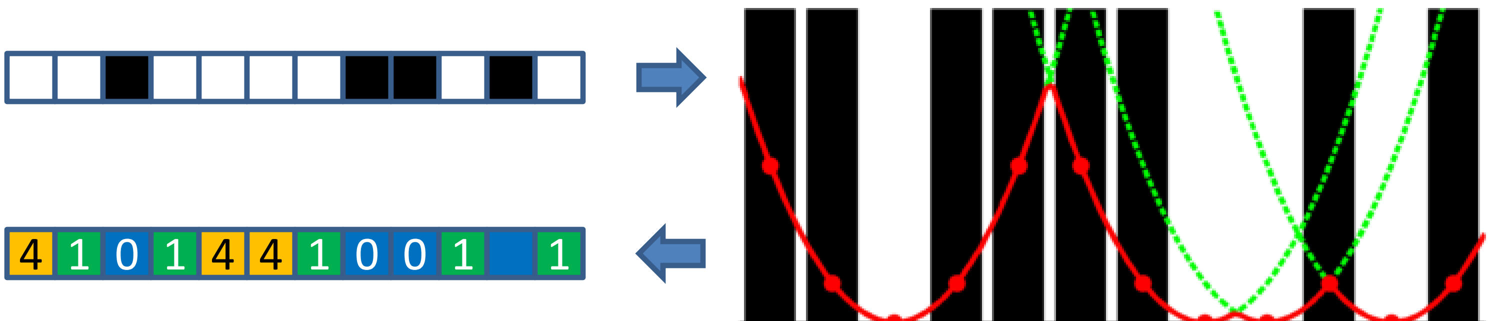


Energy on grid points with optimal observation / deformation tradeoff

(Squared) Distance Transform

- Minimum of upward paraboloids extending from ink pixels only, rooted at zero

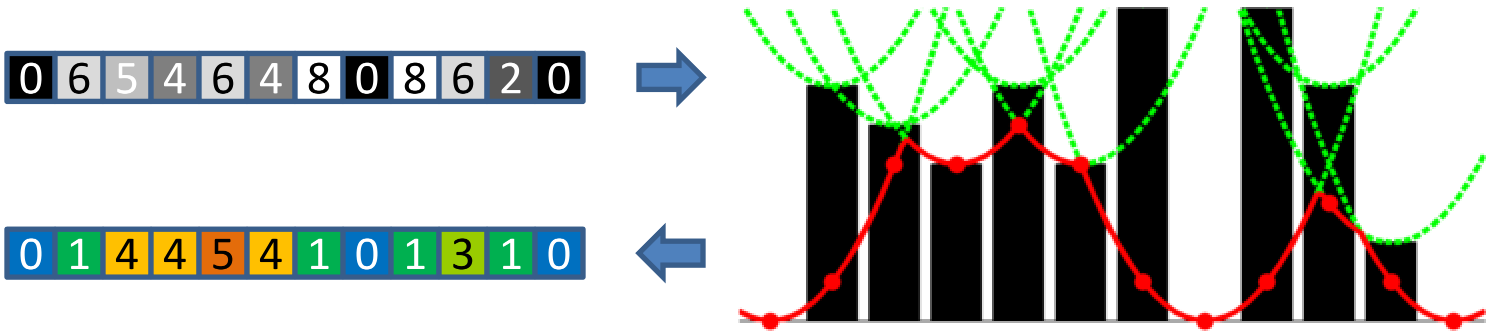
1D Example:



Note: Computational complexity grows linearly with number of pixels

Generalized Distance Transform

- Minimum of upward paraboloids *at every pixel* but rooted at pixel value
  - *Still linear complexity in number of pixels*

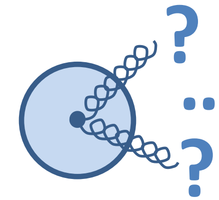


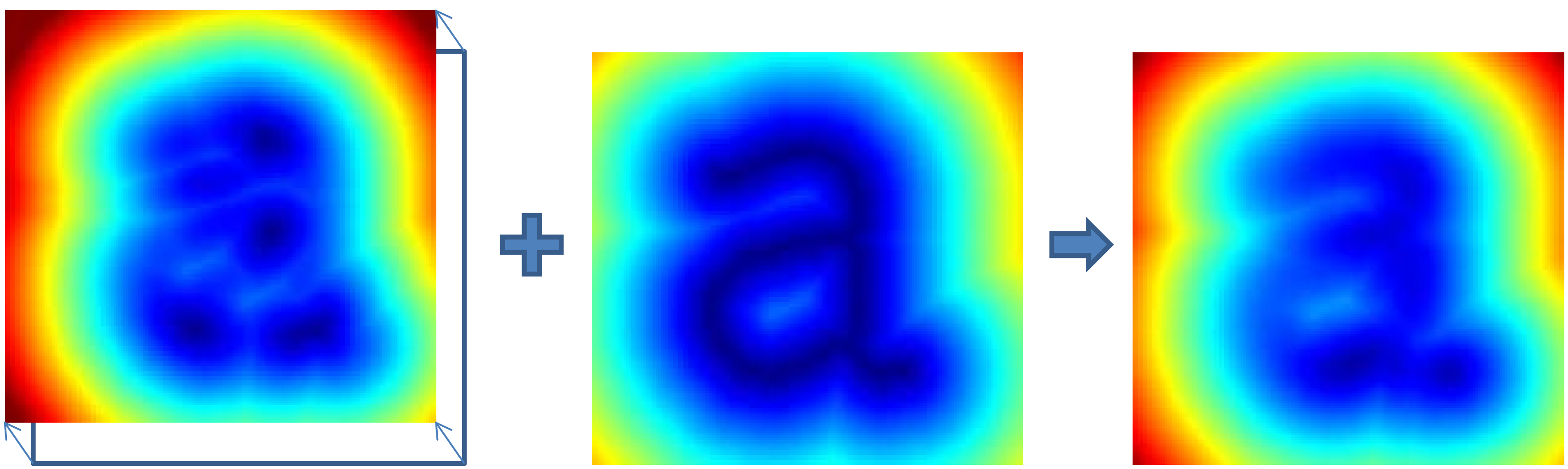
- Intuition:

  The local value can be beaten by a better one nearby

Efficient Energy Minimization

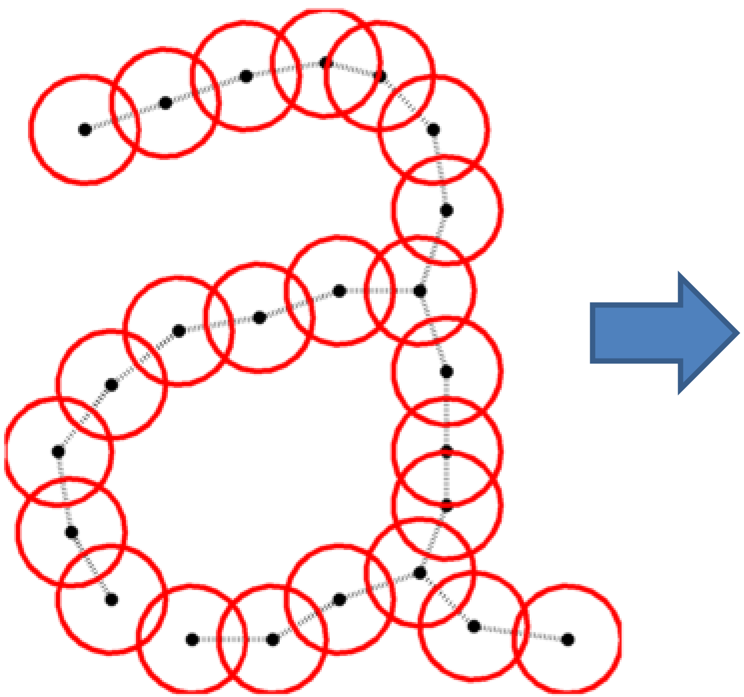
- General case: node + arbitrary structure
  - Translate energy of child structure(s) by offset
  - Apply generalized distance transform
  - Add to single-node energy

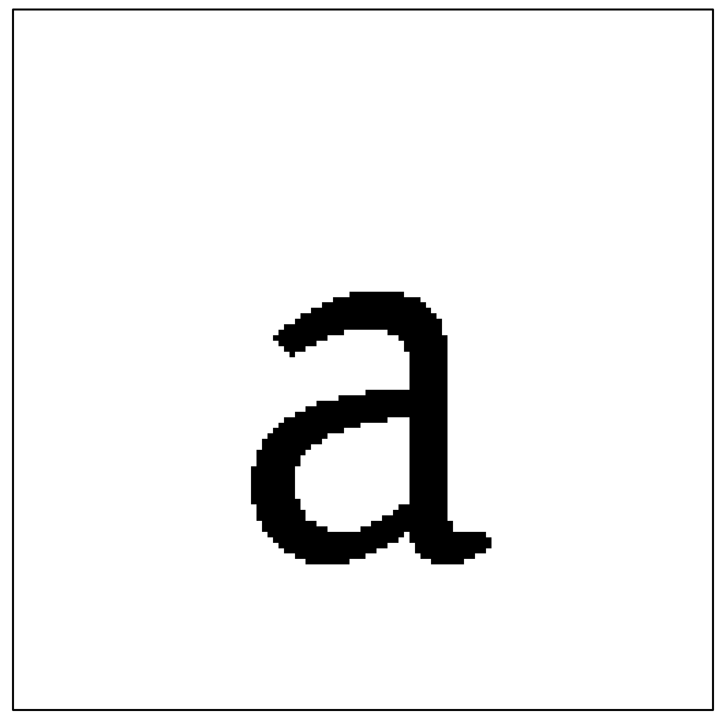




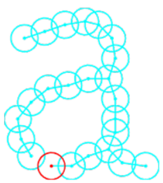
Model Matching Visualization

- Demonstration with simple example:
  Match model **a** to image

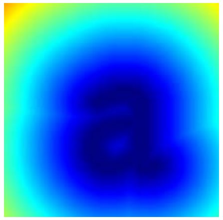




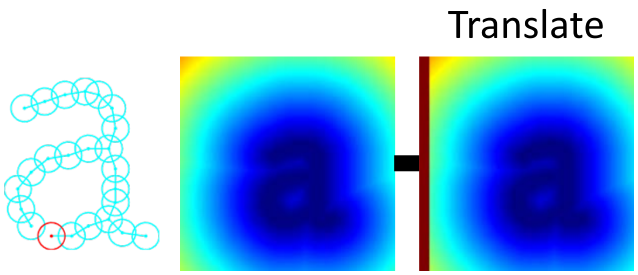
Model Matching Visualization



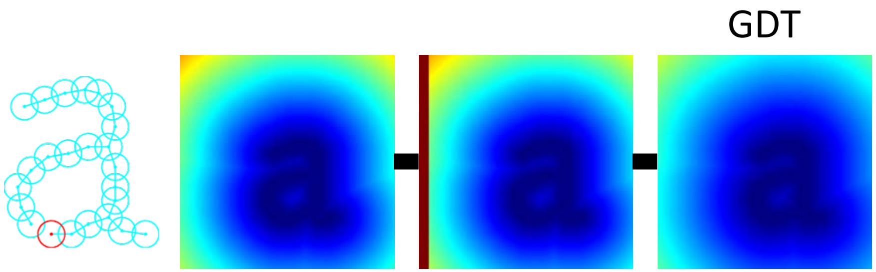
Single Node



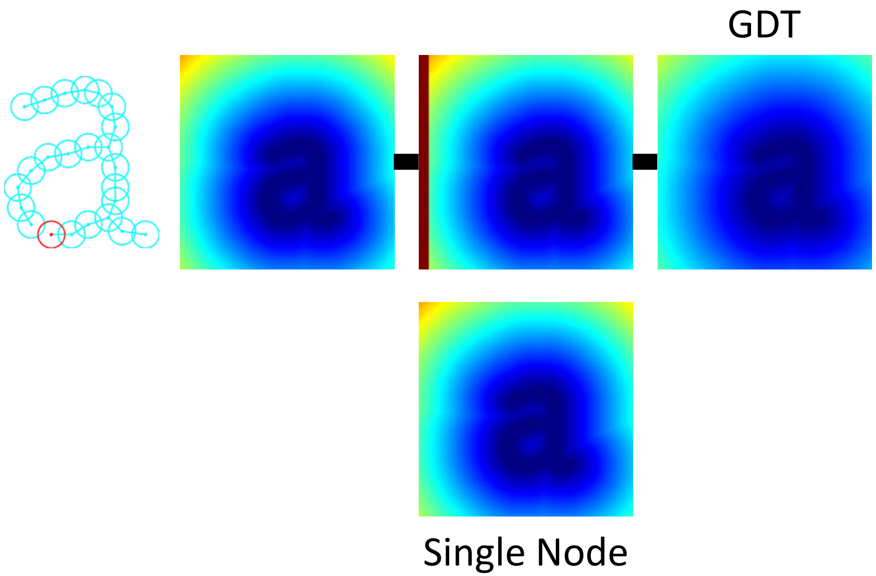
Model Matching Visualization



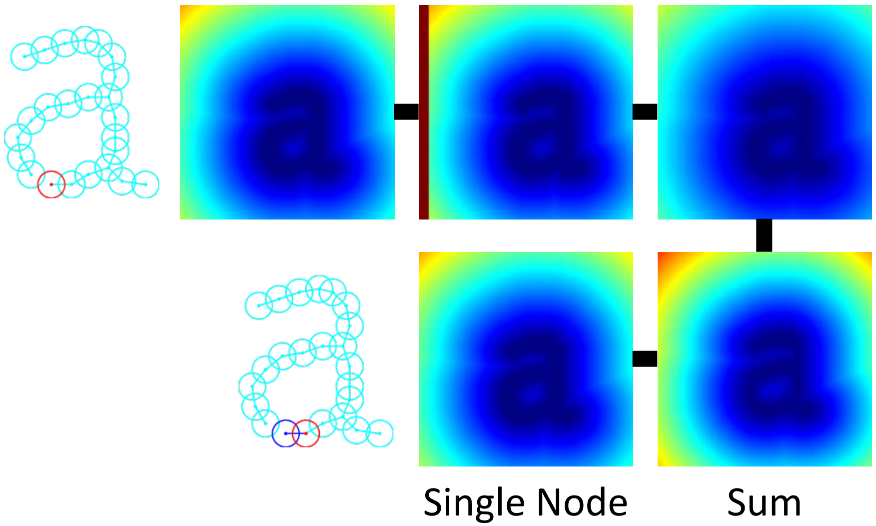
Model Matching Visualization



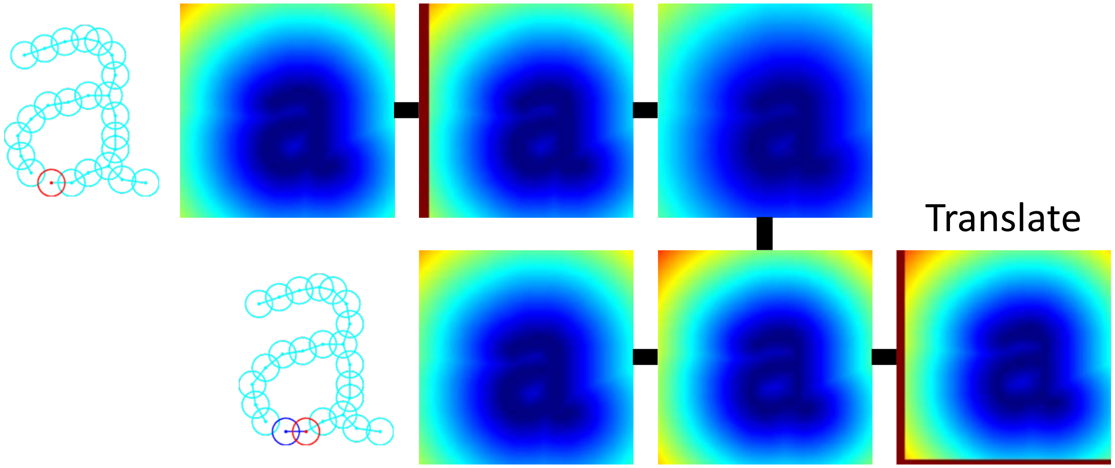
Model Matching Visualization



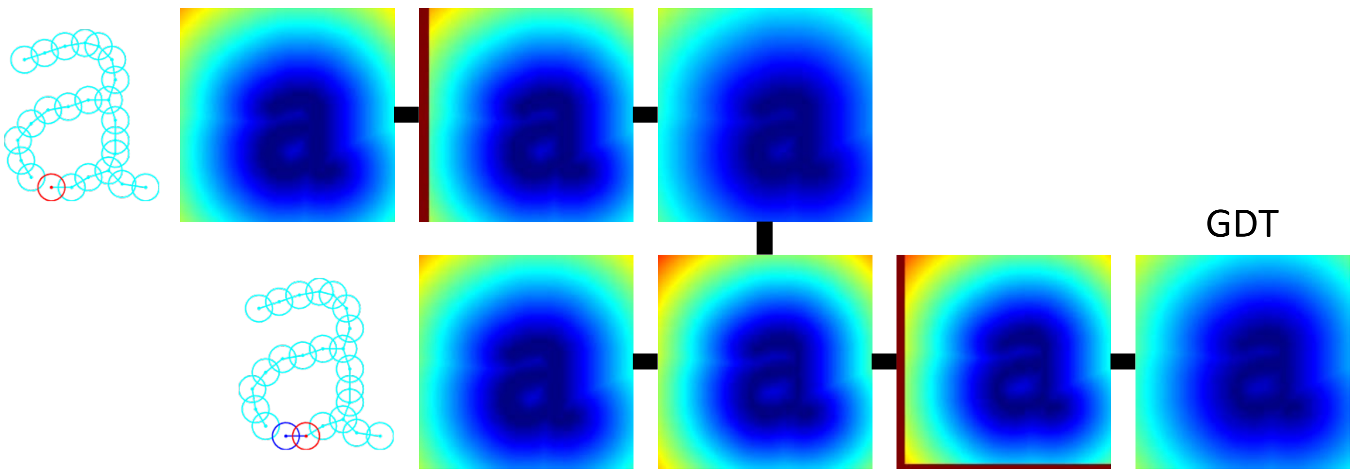
Model Matching Visualization



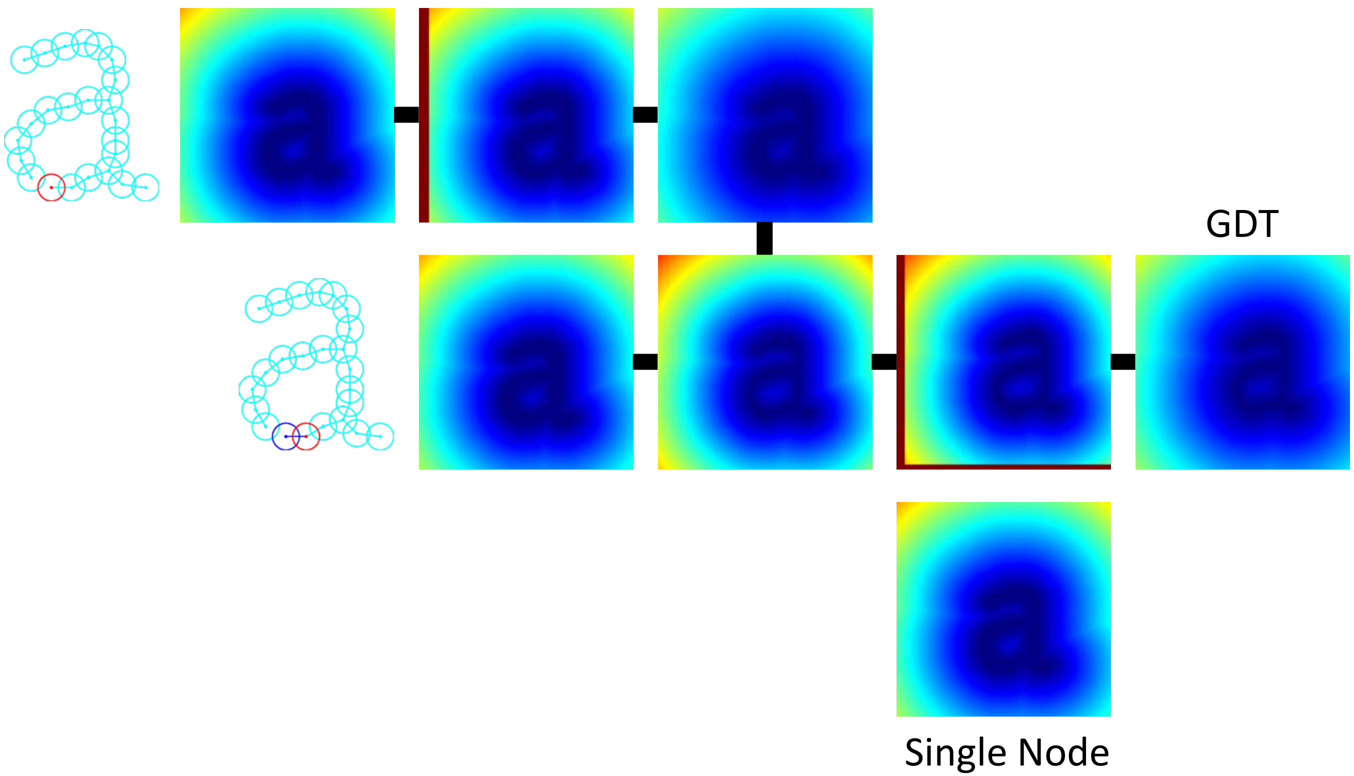
Model Matching Visualization



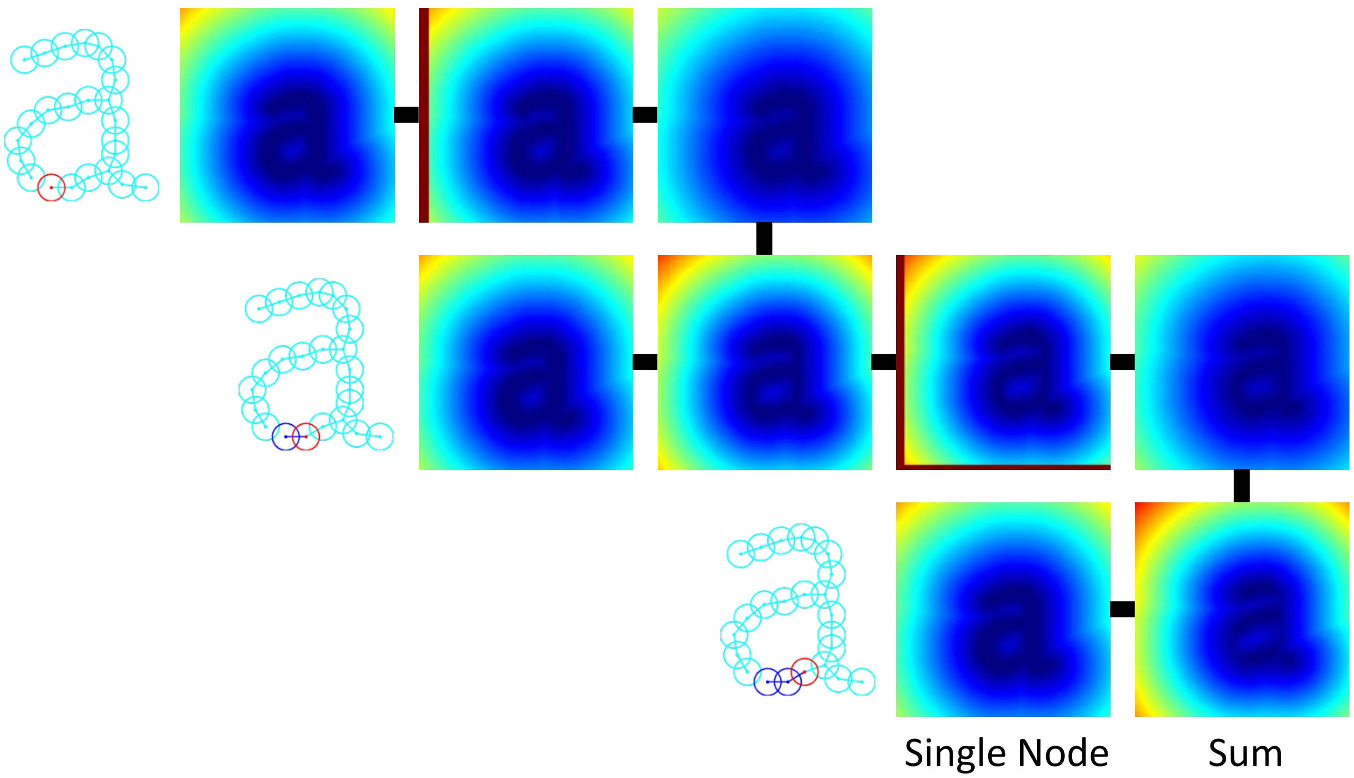
Model Matching Visualization



Model Matching Visualization

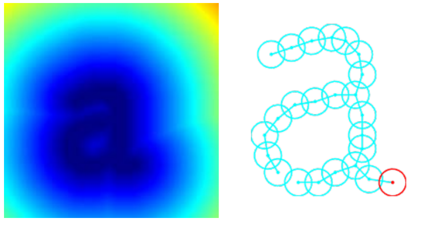


Model Matching Visualization



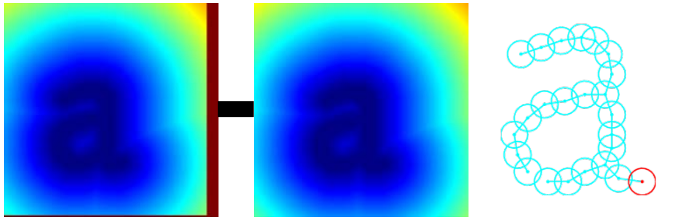
Model Matching Visualization

Single Node



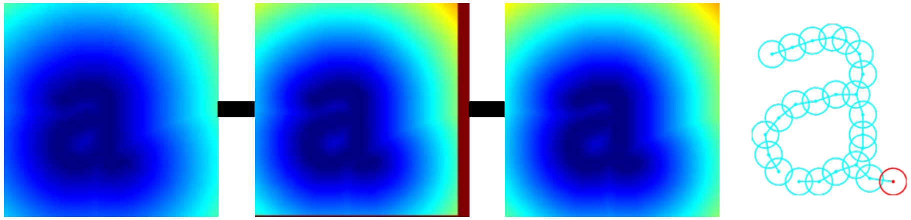
Model Matching Visualization

Translate

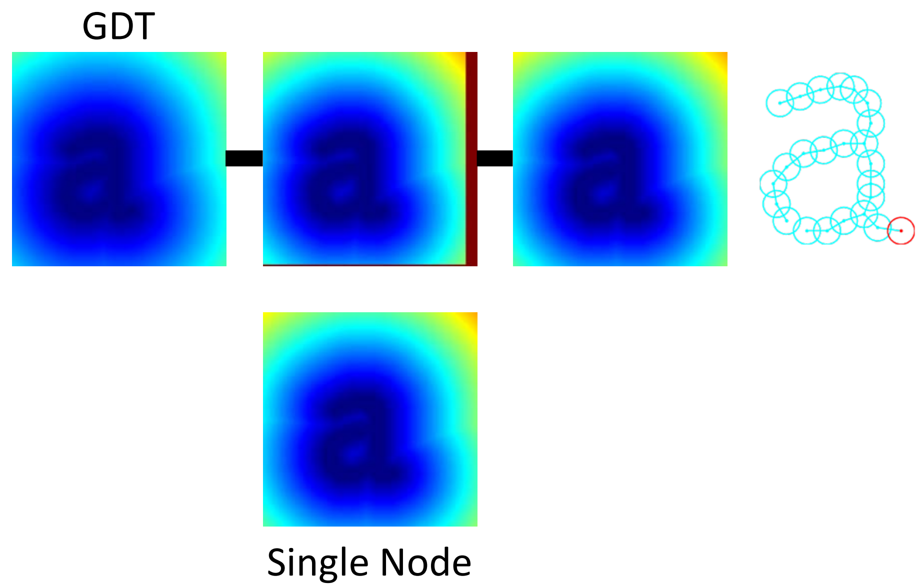


Model Matching Visualization

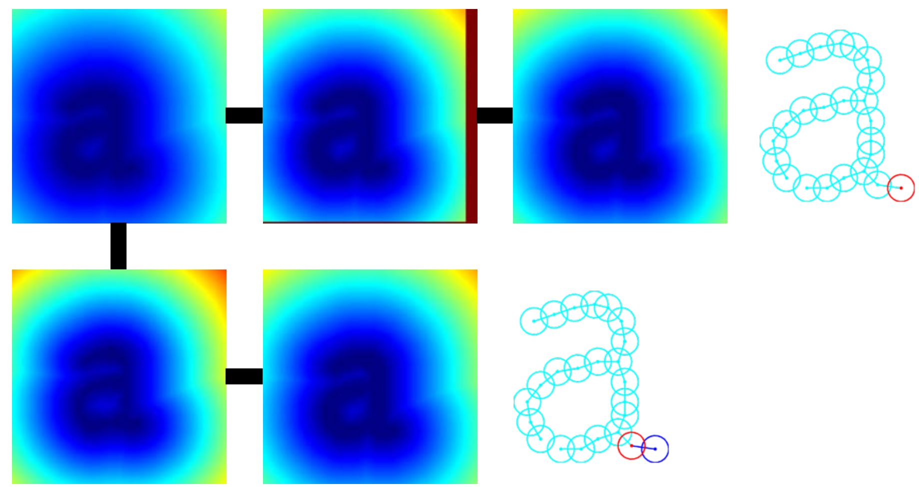
GDT



Model Matching Visualization



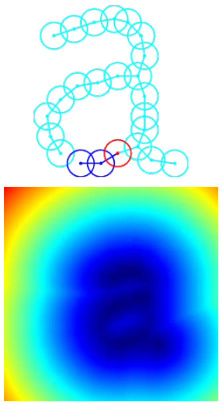
Model Matching Visualization

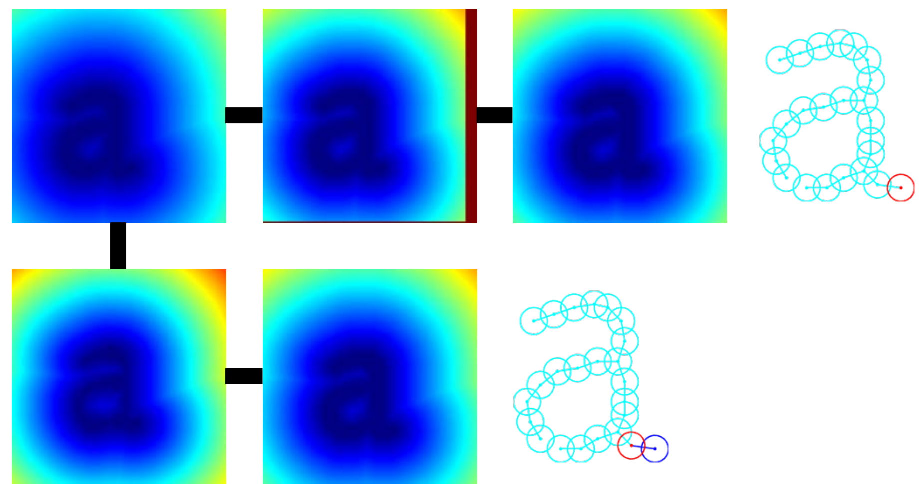


Sum

Model Matching Visualization

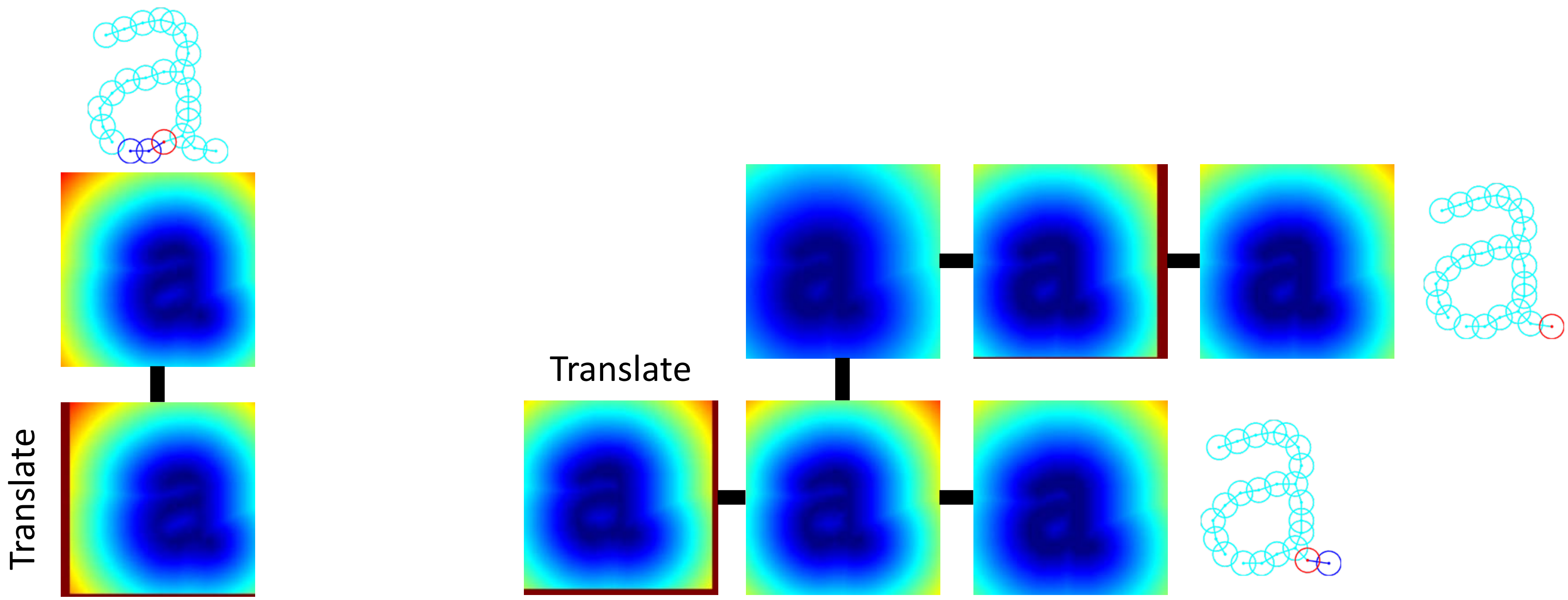
Subtree



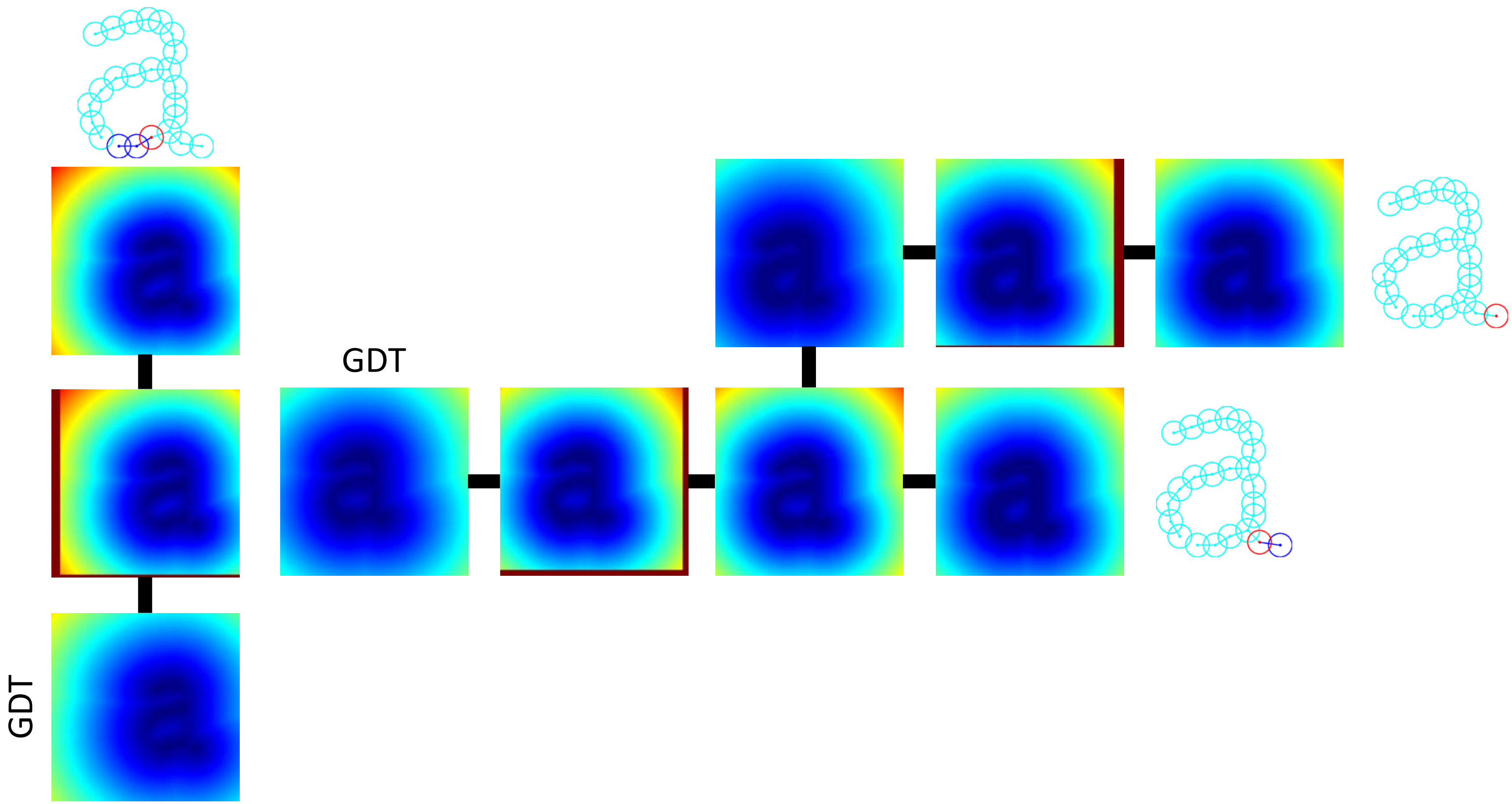


Sum

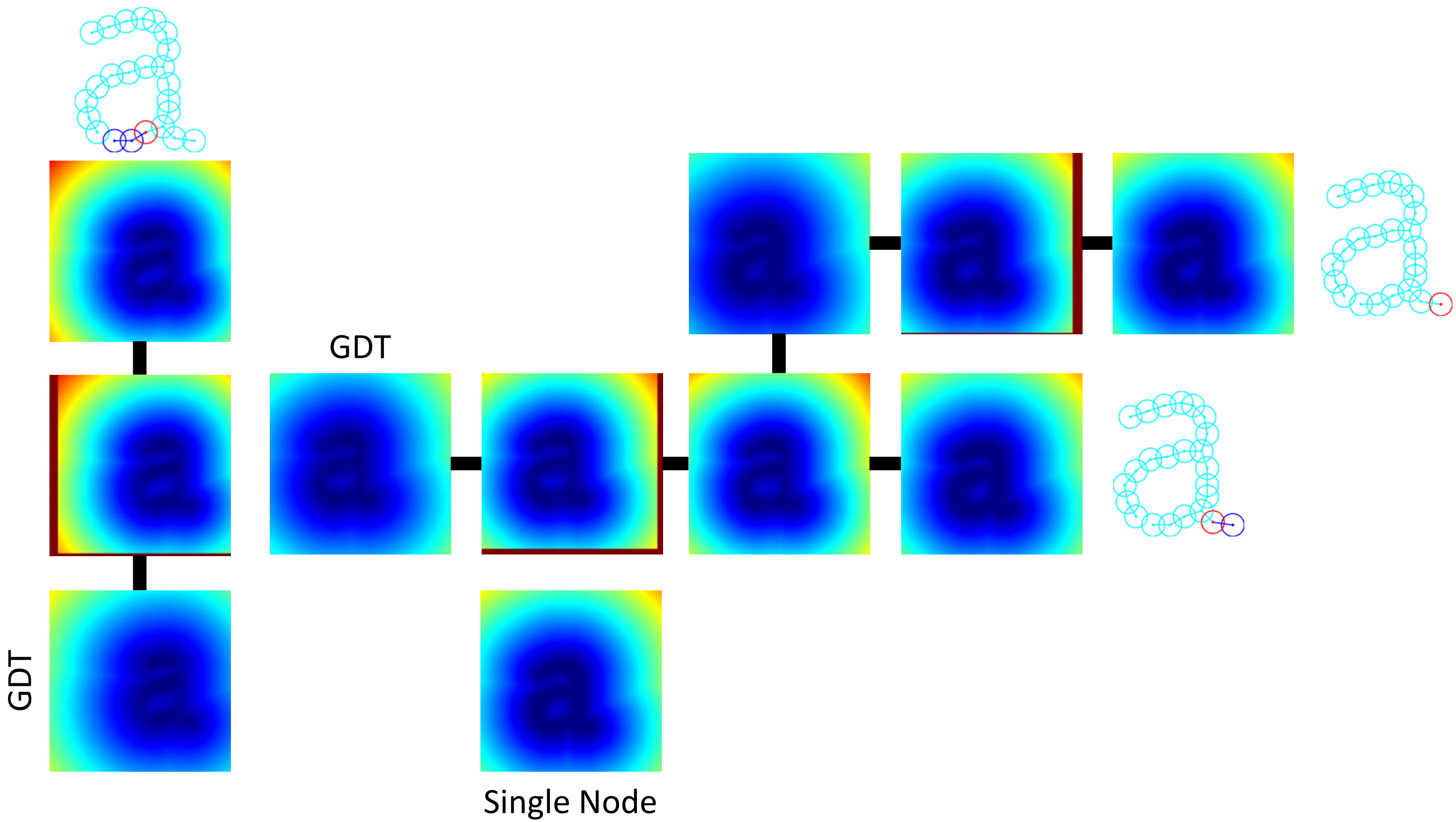
Model Matching Visualization



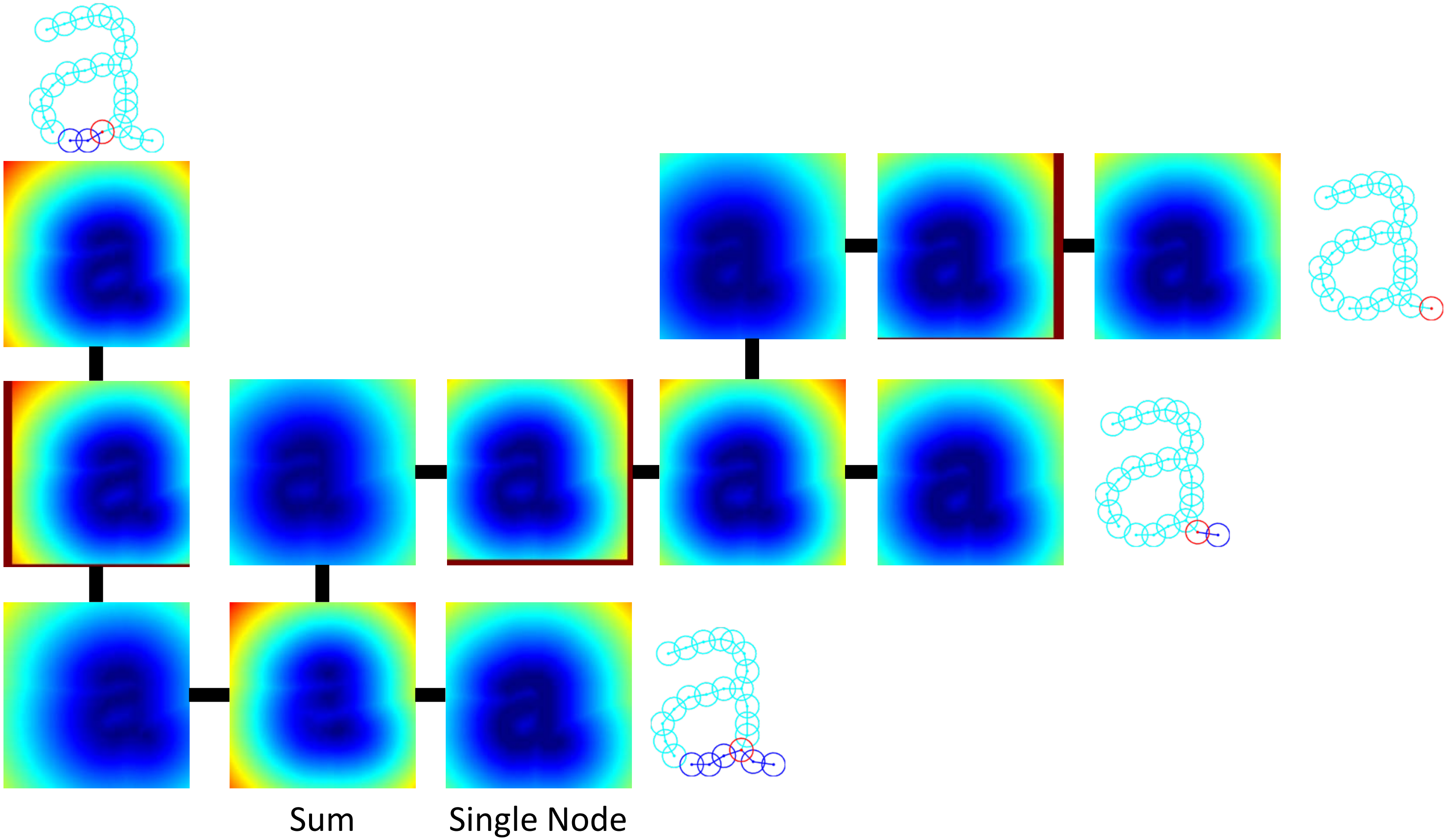
Model Matching Visualization



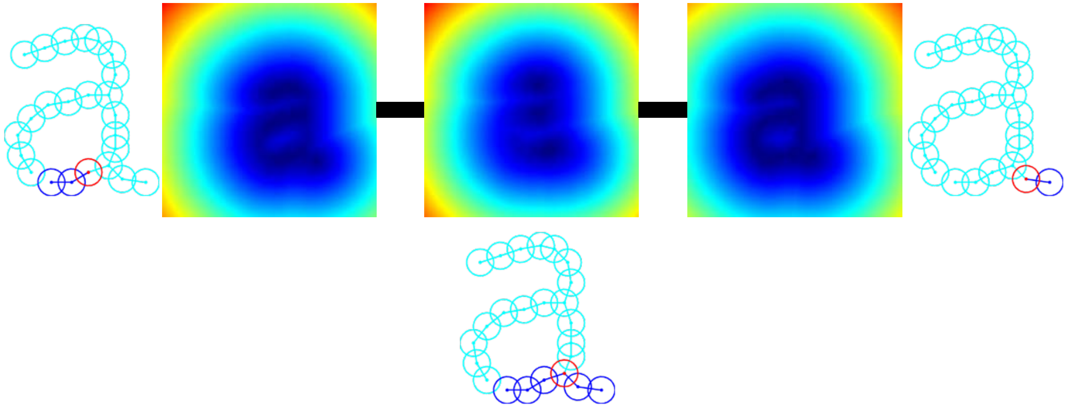
Model Matching Visualization



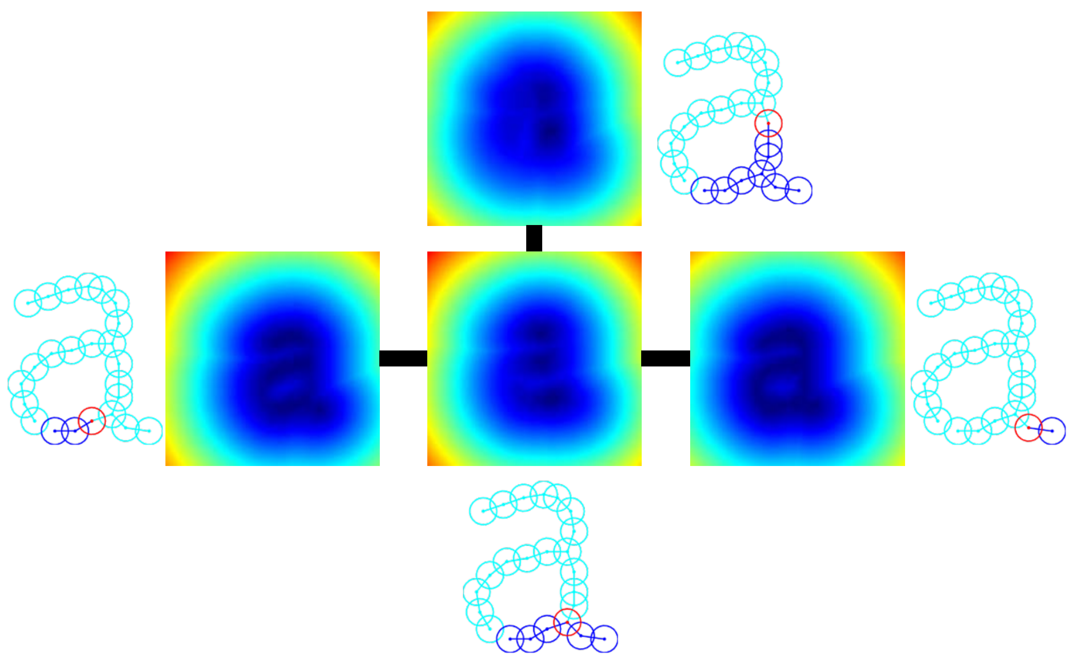
Model Matching Visualization



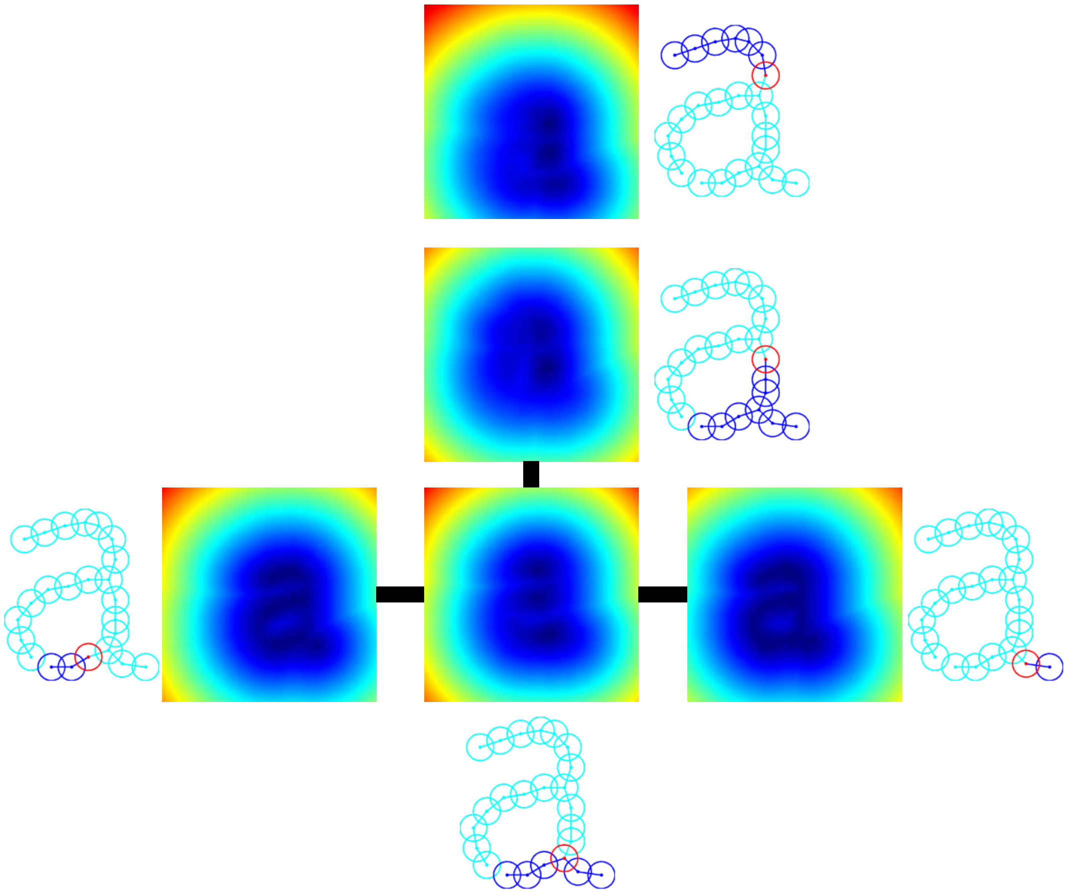
Model Matching Visualization



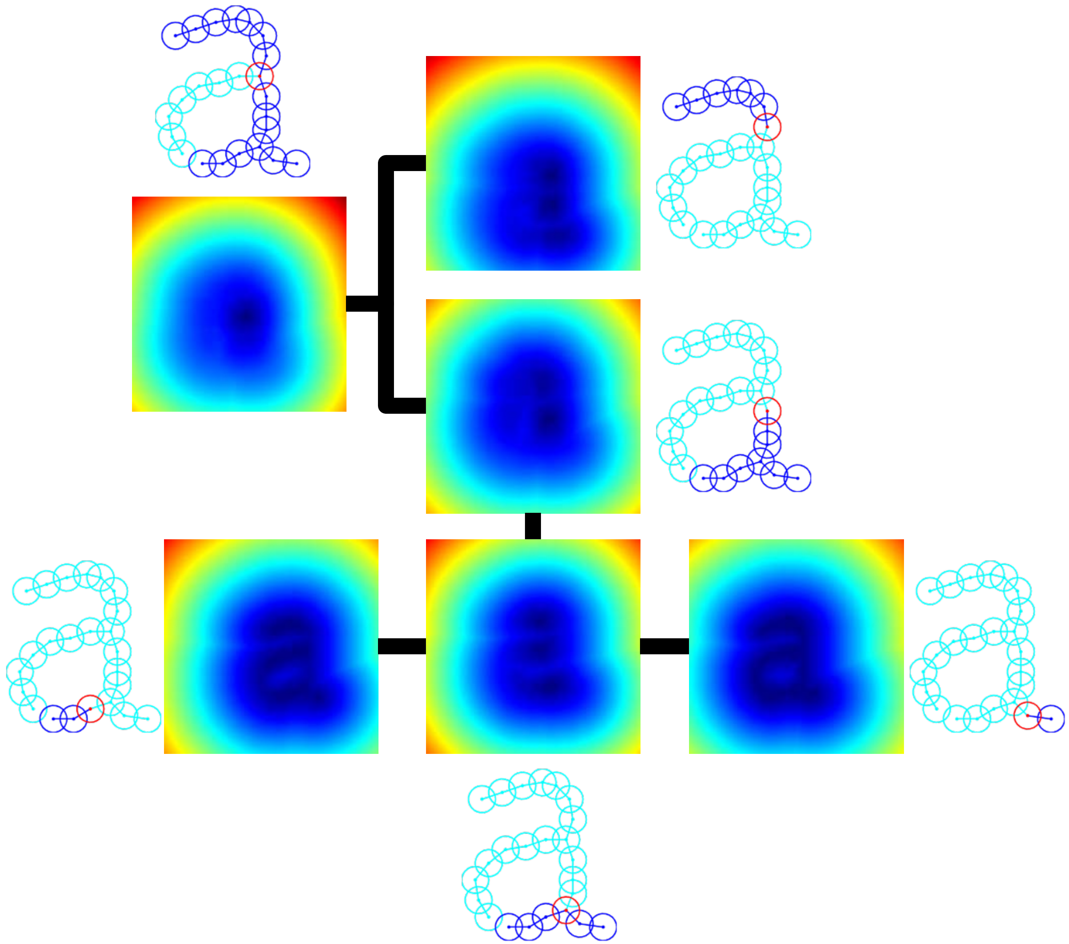
Model Matching Visualization



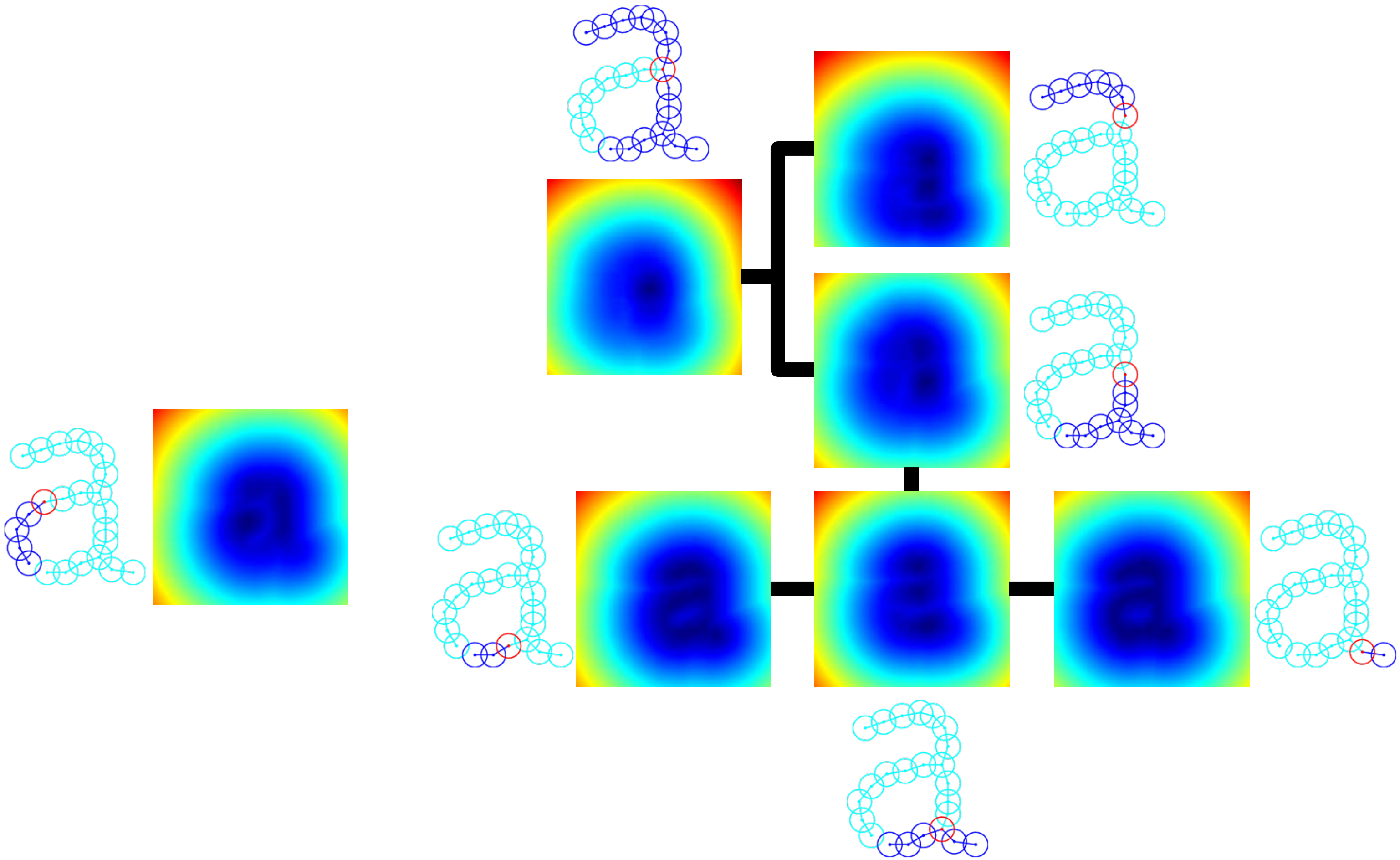
Model Matching Visualization



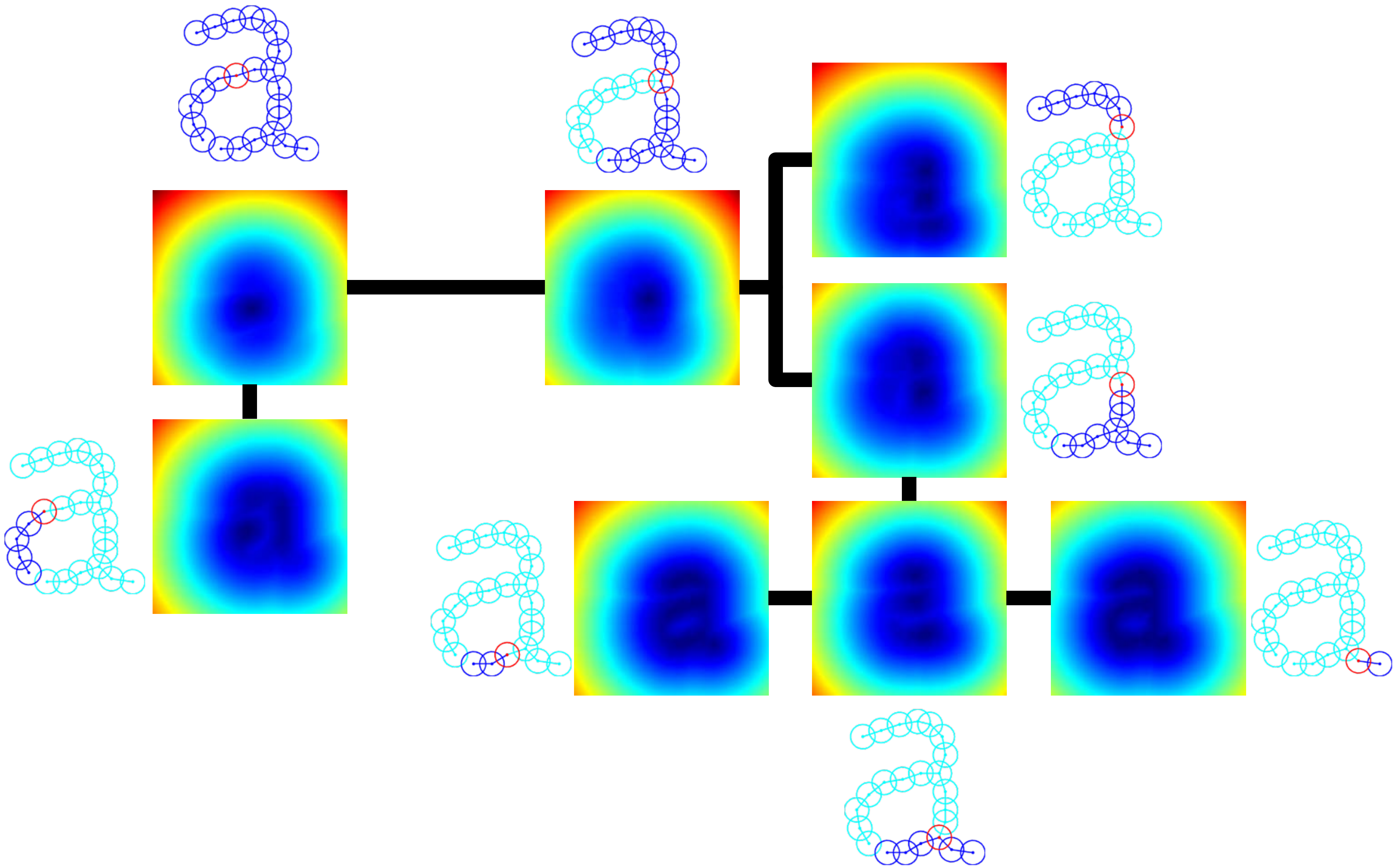
Model Matching Visualization



Model Matching Visualization

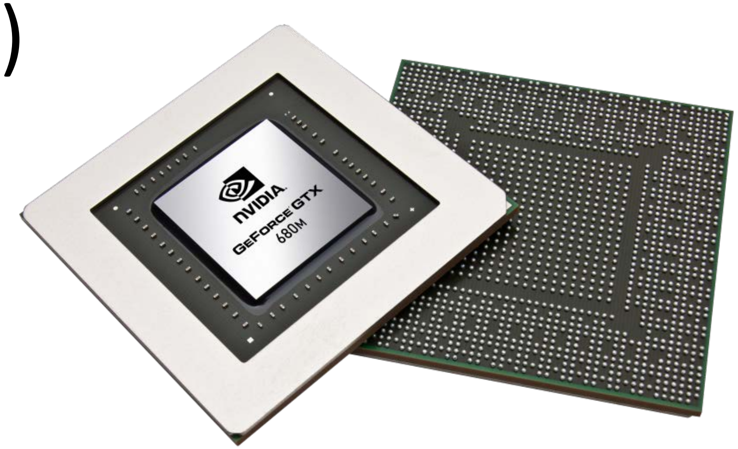


Model Matching Visualization



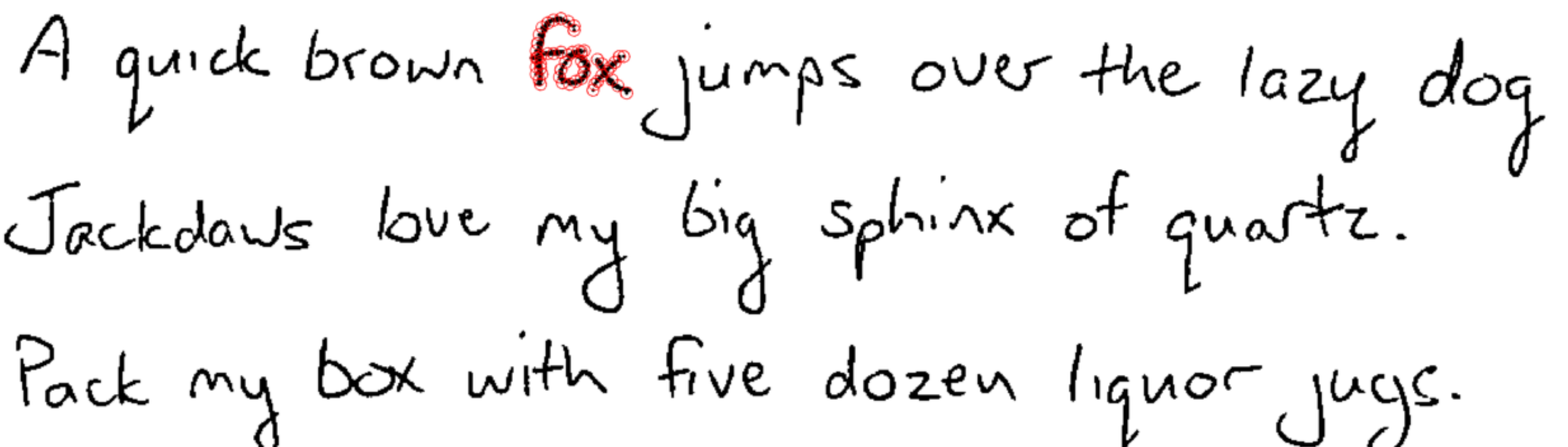
Parallel GDT

- Optimum model fit requires:
  - One translation per node
  - One GDT per node
- Work scales with number of image pixels
- Fast parallel computation on graphics processing unit (GPU)



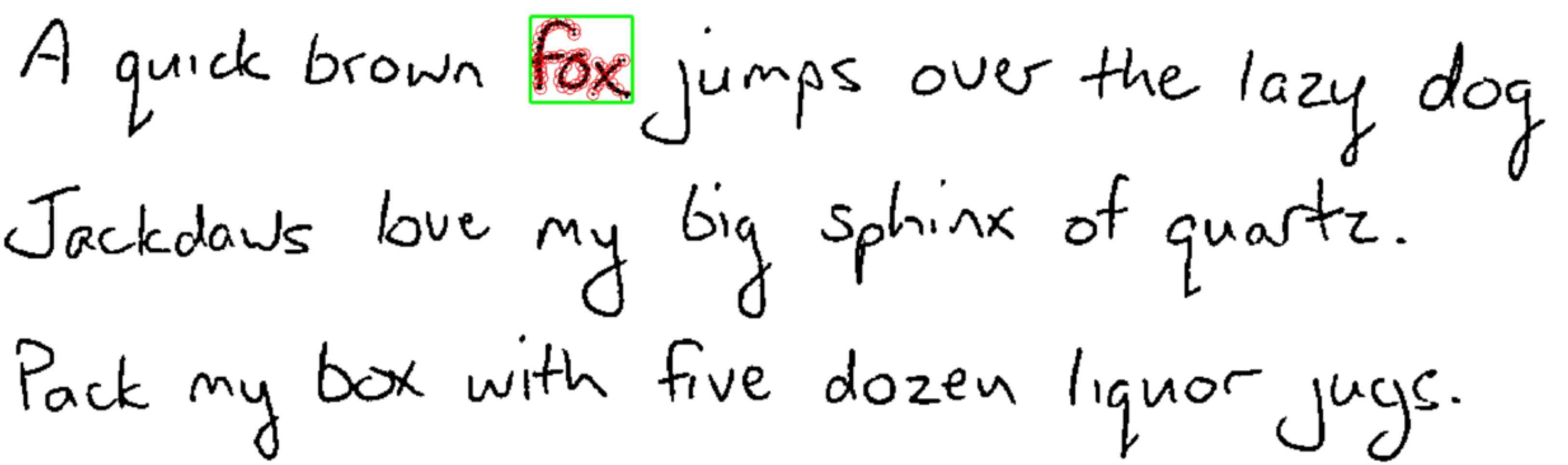
Configuration Recovery

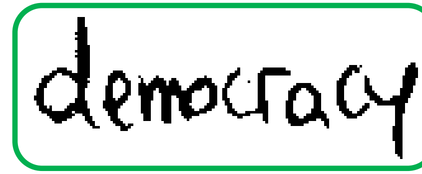
- Energy optimization/model matching is just big dynamic programming problem
- Trace back DP winner to recover configuration
- Useful for display/localization

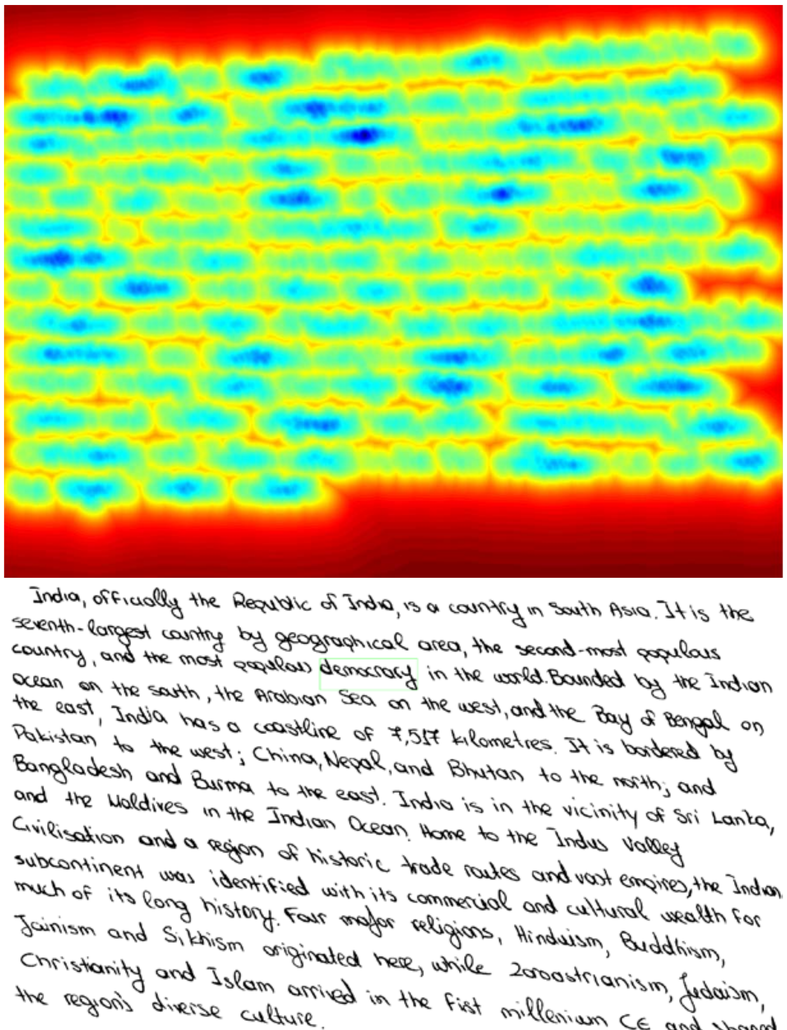


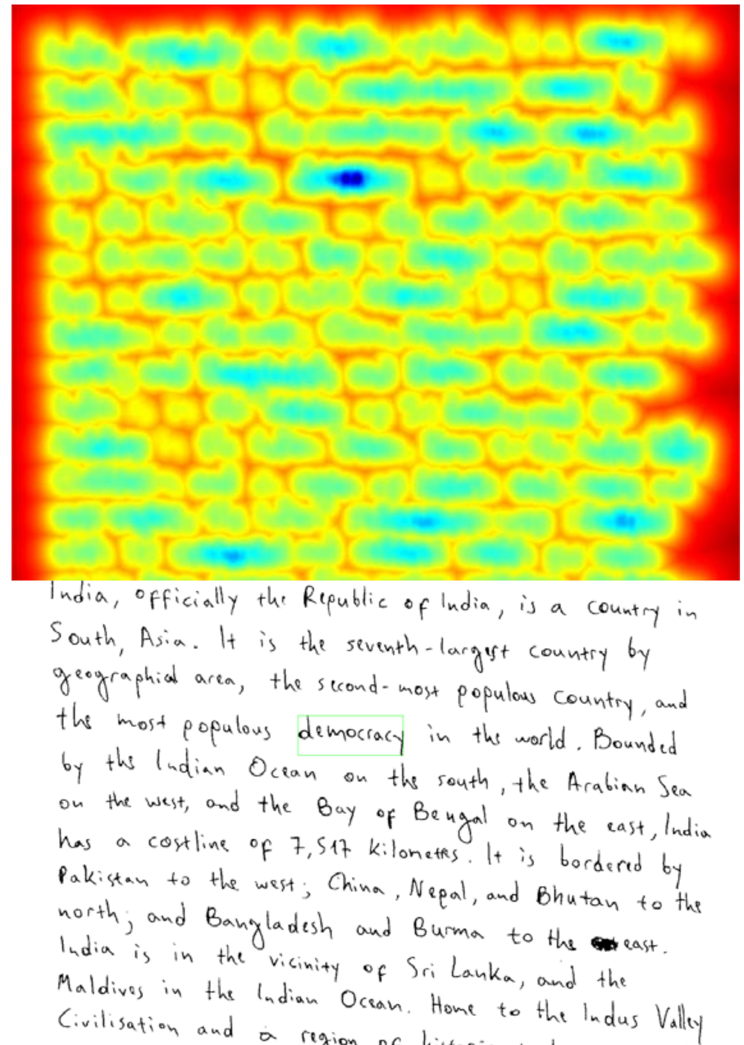
Configuration Recovery

- Energy optimization/model matching is just big dynamic programming problem
- Trace back DP winner to recover configuration
- Useful for display/localization



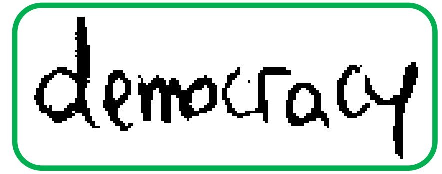
Sample Result: Query = 

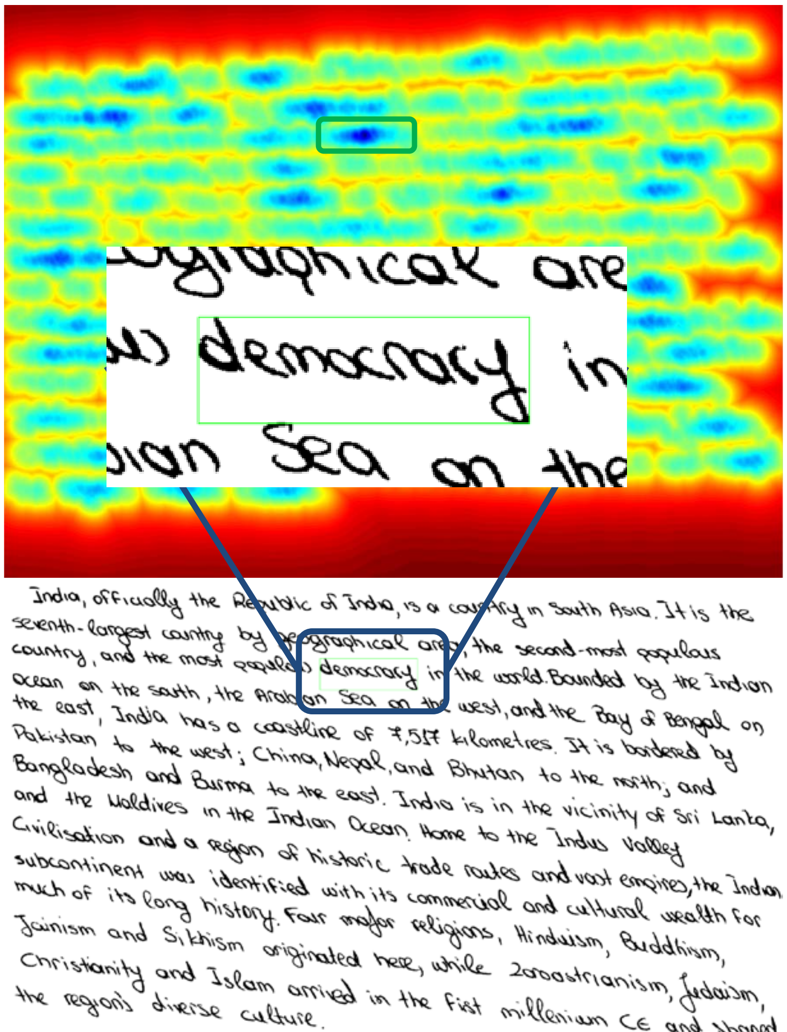


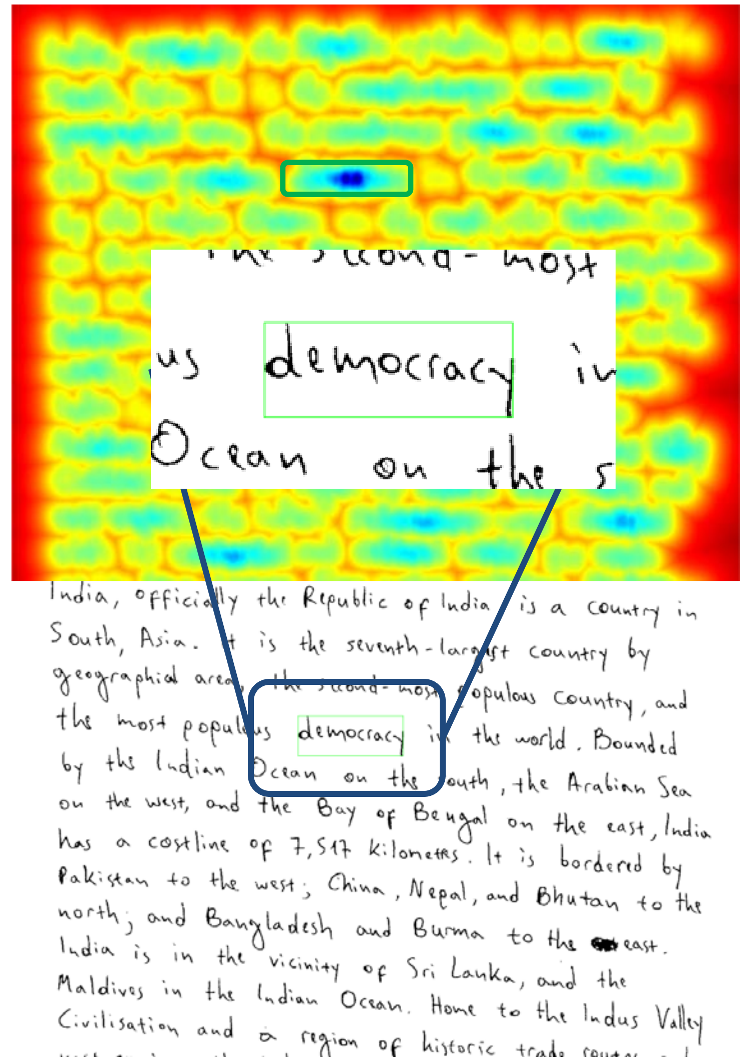


Note: left/right color scales do not match.

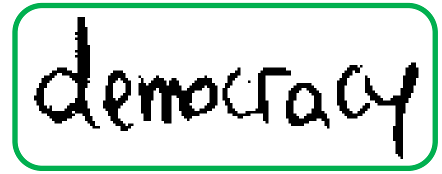
Sample Result: Query =

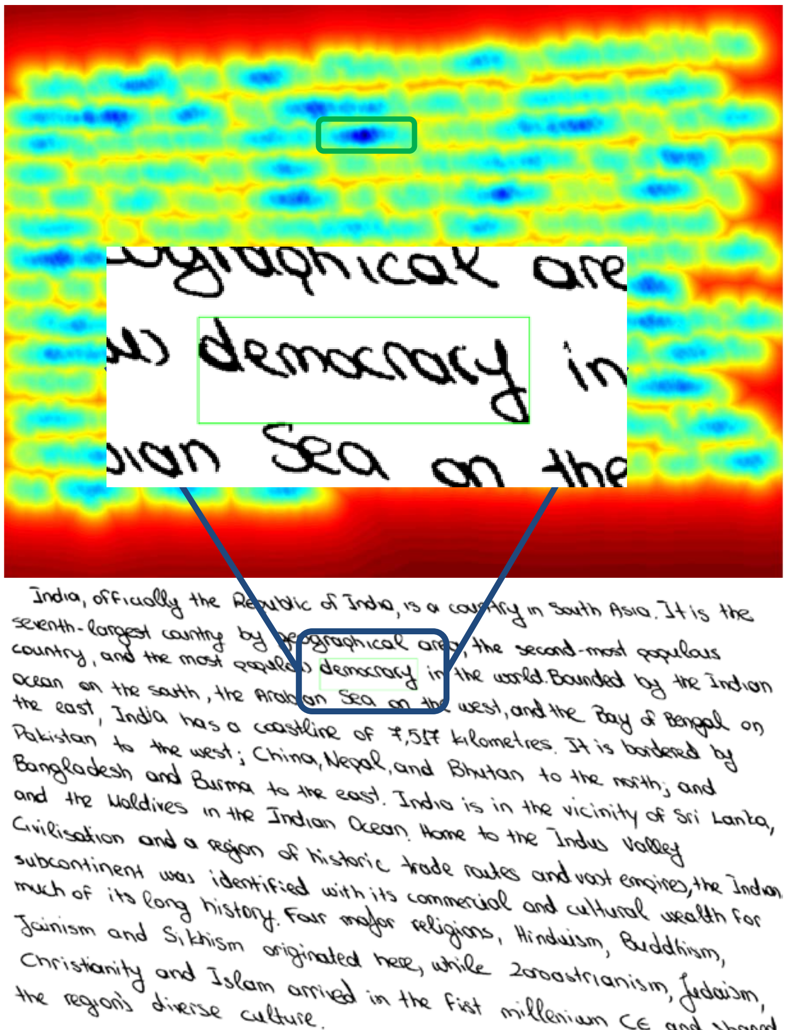


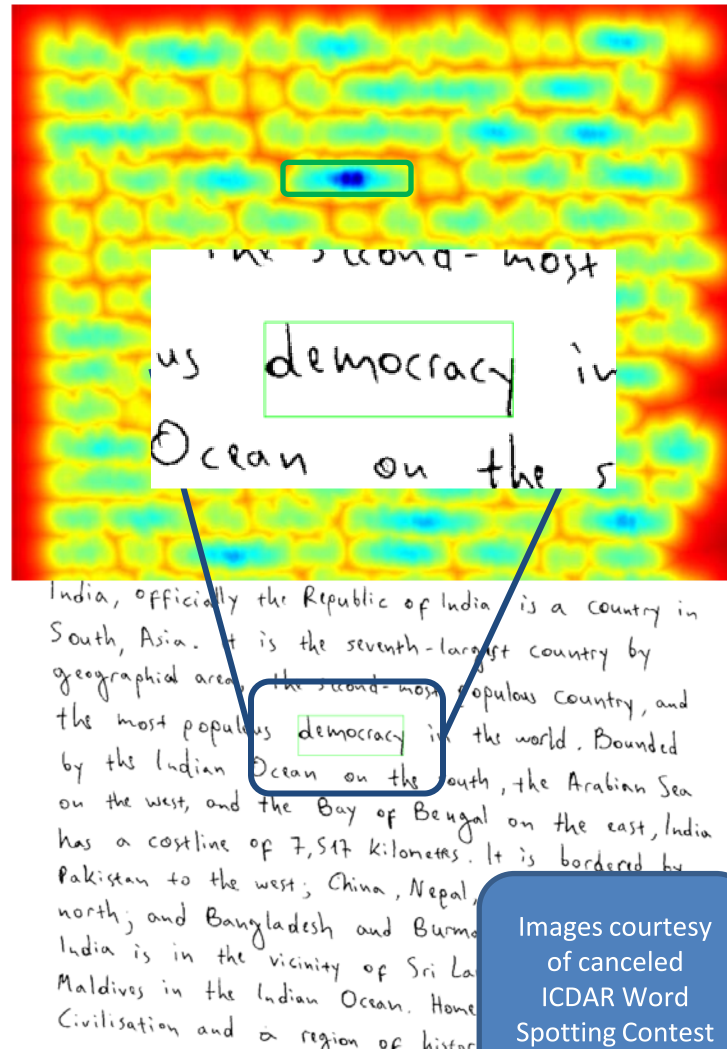




Note: left/right color scales do not match.

Sample Result: Query = 

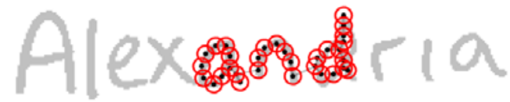
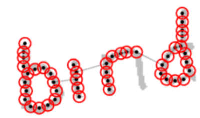


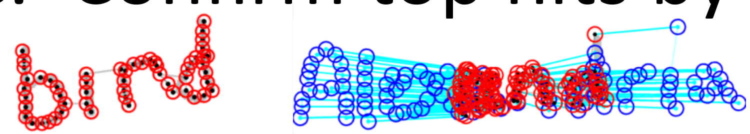


Images courtesy of canceled ICDAR Word Spotting Contest

Note: left/right color scales do not match.

Match Confirmation

- Model matches ink, ignores noise/context
  - Will match **and** to **Alexandria**: 
  - Will match **bird** to **bind**: 
- Whitespace not considered in model
- Expedient heuristic: Confirm top hits by reverse match

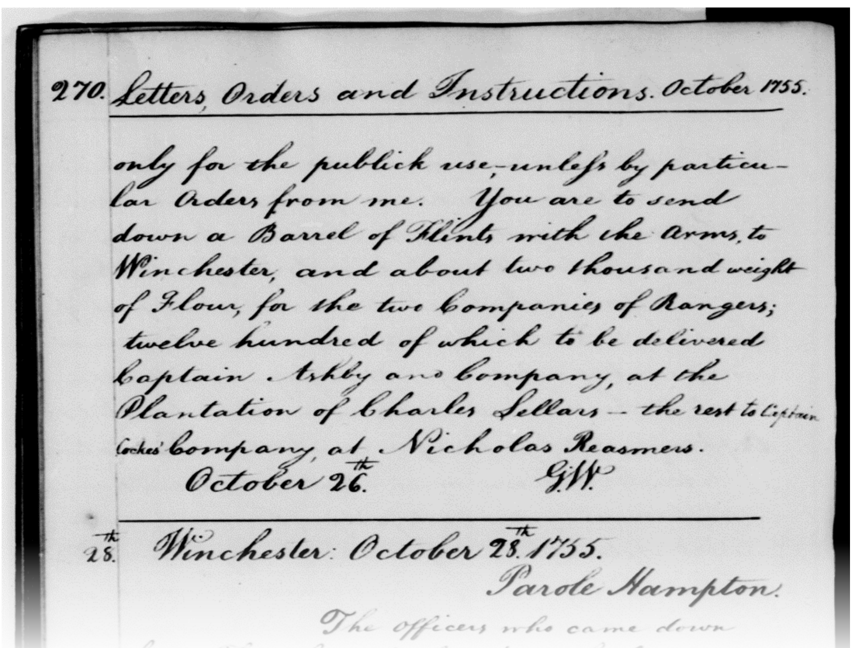


  - Build model of target area & match to query
  - Match energy is greater of the two directions (scaled by number of nodes)

Experimental Data Sets

George Washington (GW20)

- 20 pages; 4685 words
- English cursive script



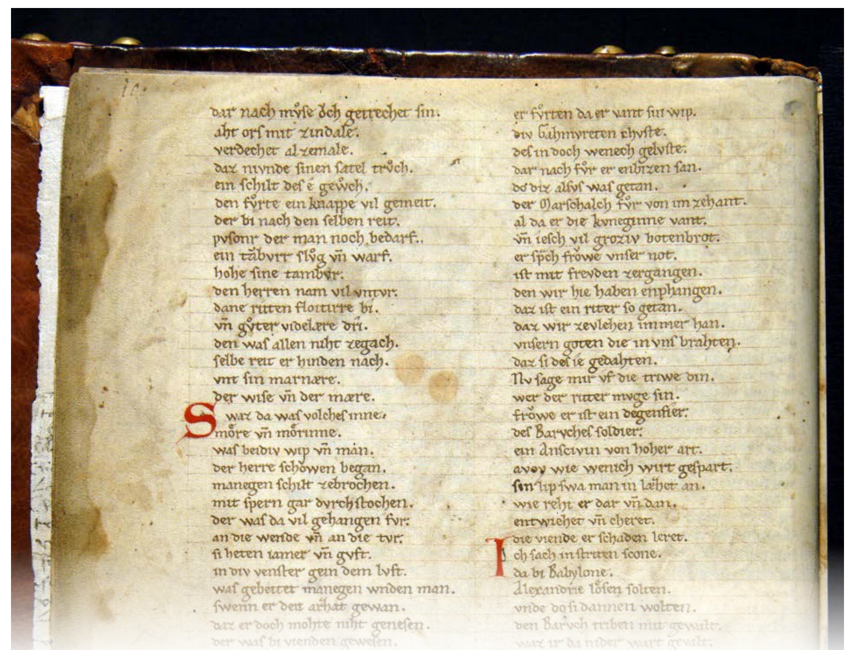
270. Letters, Orders and Instructions. October 1755.

only for the publick use,—unless by particu-
lar Orders from me. You are to send
down a Barrel of Flints with the Arms, to
Winchester, and about two thousand weight
of Flour, for the two Companies of Rangers;
twelve hundred of which to be delivered
Captain Ashby and Company, at the
Plantation of Charles Sellars — the rest to Captain
Cocke's Company, at Nicholas Reasmers.
October 26th. G.W.

28th. Winchester: October 28th 1755.
Parole Hampton.
The officers who came down

Parzival

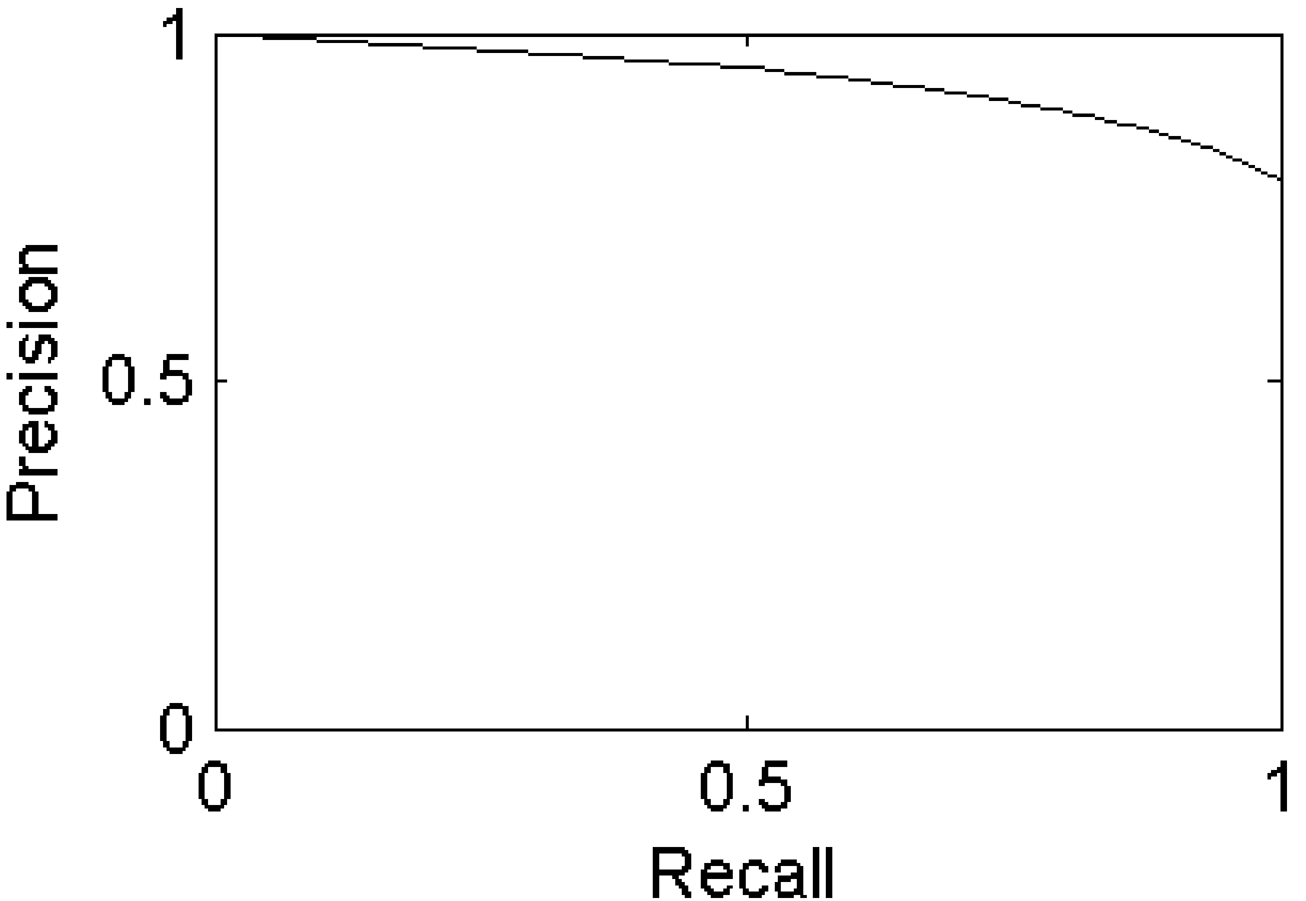
- 47 pages; 18,918 words
- German medieval lettering



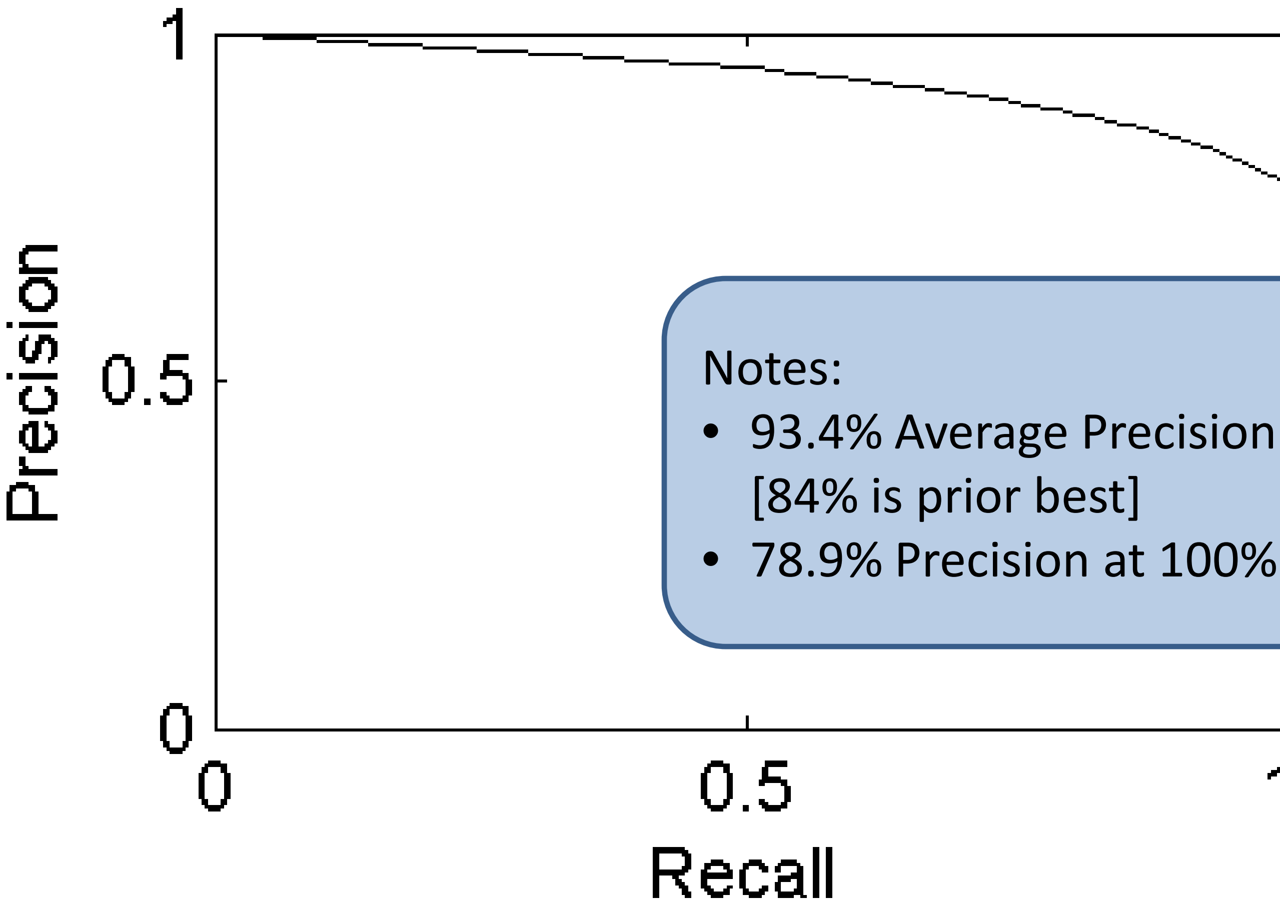
Methodology

- Used train/test split from Frinken et al. [PAMI'12]
- Each non-stopword in training set is a query
  - Some appear multiple times in training set
  - Run retrieval on all instances & take high scores
- Reverse match uses segmented words
- Recall-Precision curves averaged for all queries
  - Threshold may vary from query to query
  - Cross-query calibration still requires research

George Washington

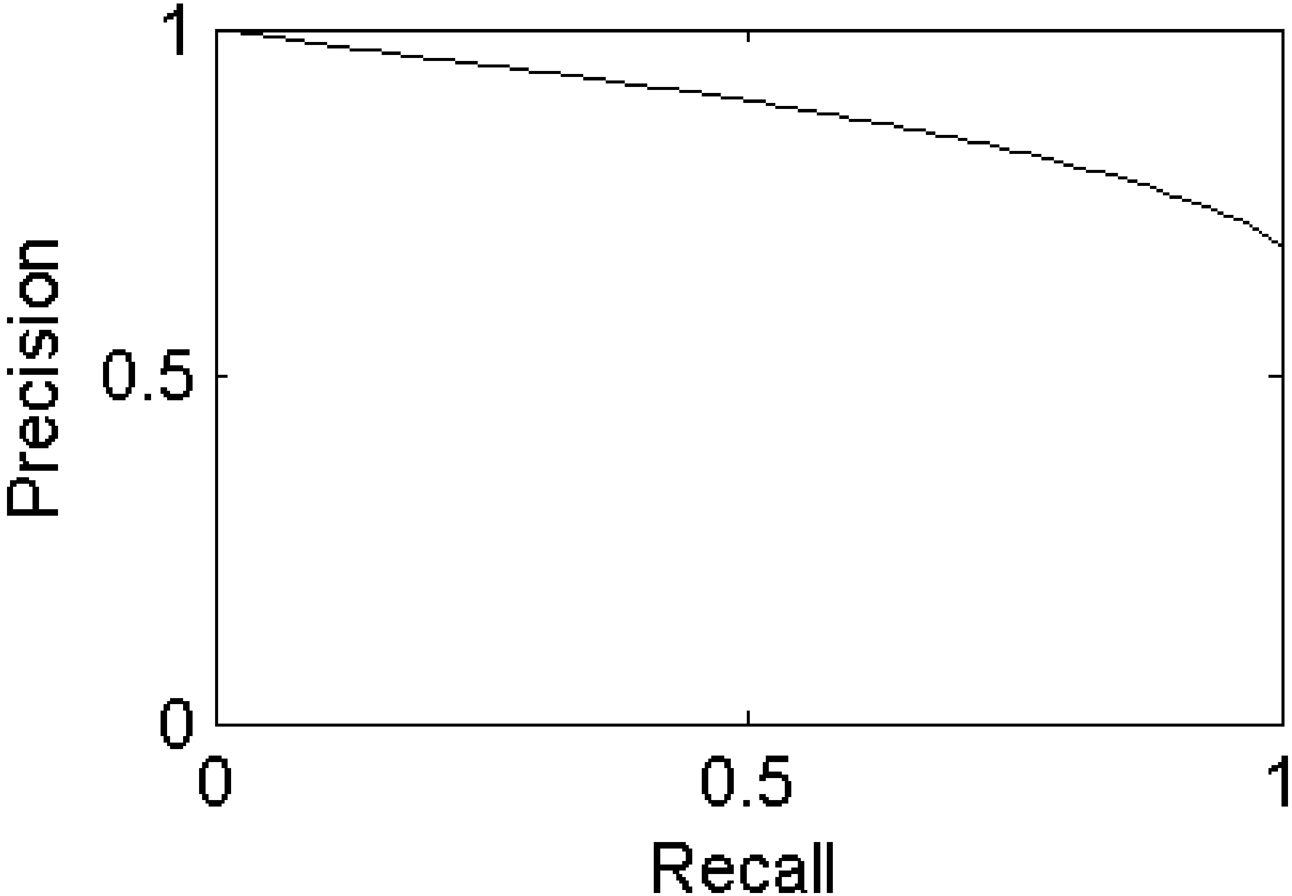


George Washington

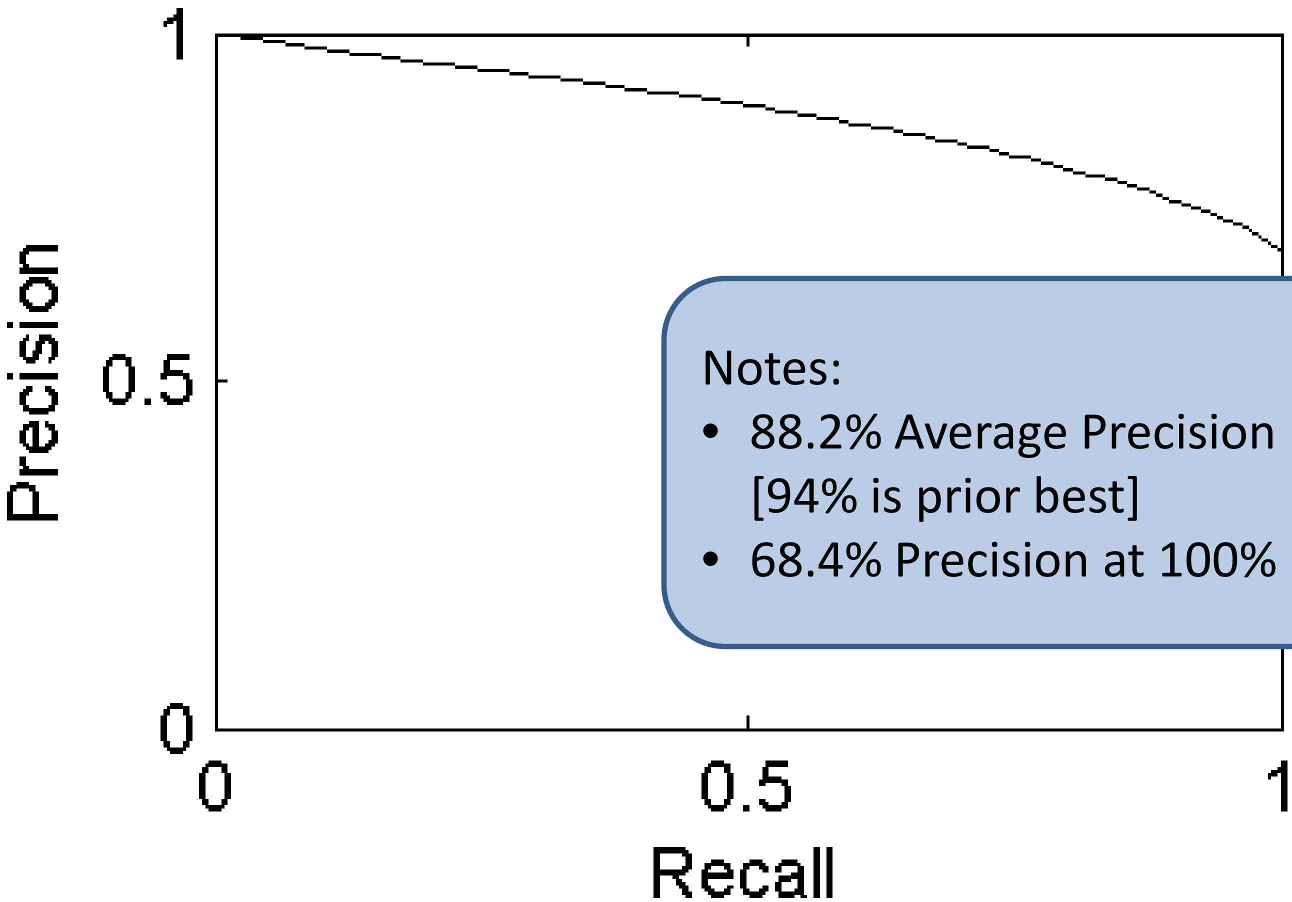


Notes:
- 93.4% Average Precision [84% is prior best]
- 78.9% Precision at 100% Recall

Parzival

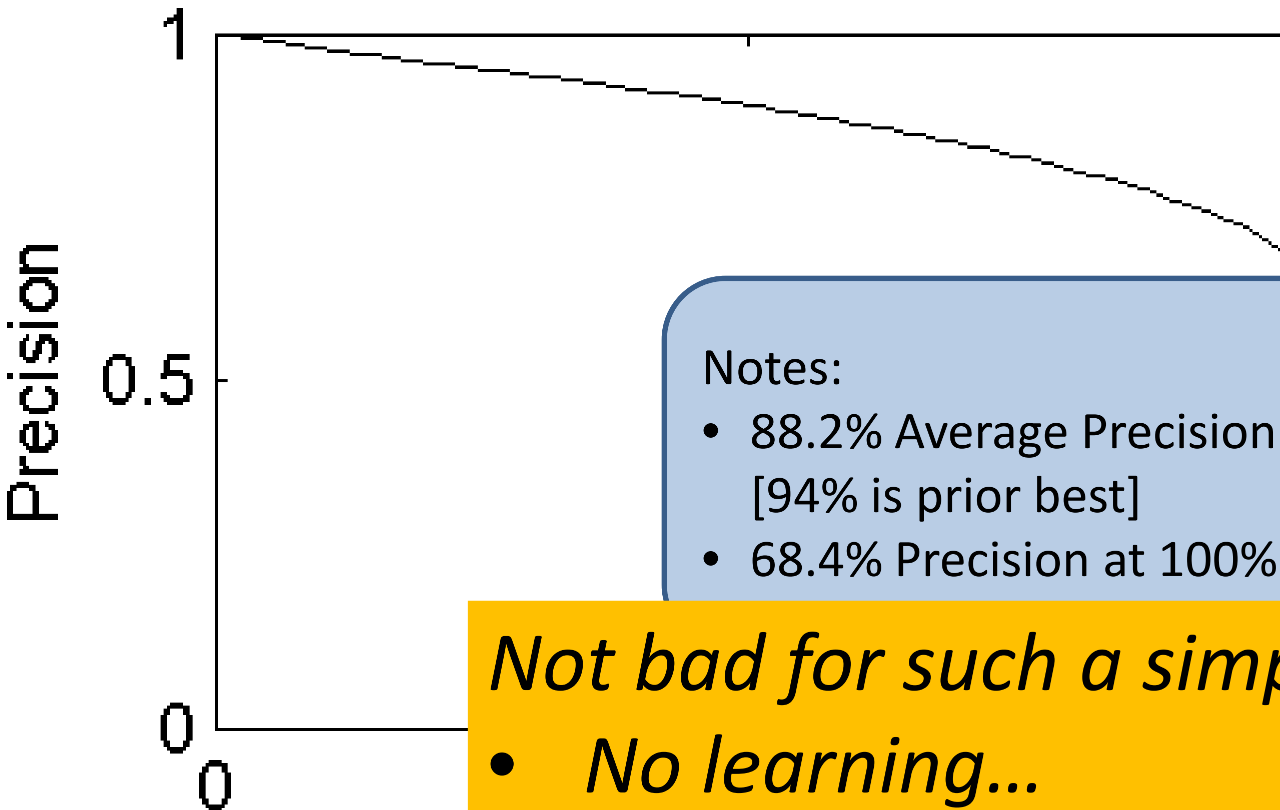


Parzival



Notes:
- 88.2% Average Precision [94% is prior best]
- 68.4% Precision at 100% Recall

Parzival



Notes:
- 88.2% Average Precision [94% is prior best]
- 68.4% Precision at 100% Recall

Not bad for such a simple model!
- *No learning…*
- *No language model…*

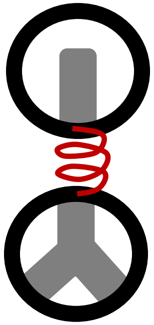
…different yet still good.

Caveat Lector

- Some dependence on handwriting style
  - Intrinsic letter forms can vary
  - Cross-style spotting requires more research
- Limited invariance to scale & rotation
  - Match model scale to text in document
  - Correct skew/rotation prior to spotting

- Speed not yet real-time for large collections
  - Roughly 2 Mpixel/second for most words

Part-Structured Promise



- Powerful matching/retrieval tool
  - Part models could be more complex
- Requires no training, language modeling, etc.
  - Easily applied to new languages, figures, etc.

- Intuitive pixel-level correspondences
  - Starting point for further processing?
- Reference code on my web page
  - I welcome opportunities to collaborate!

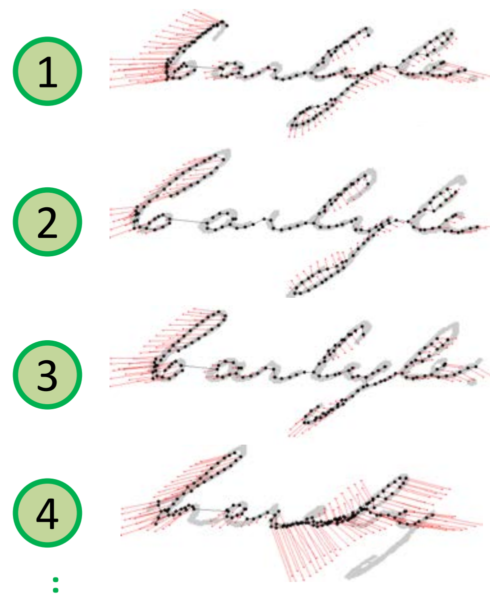
http://cs.smith.edu/~nhowe/research/code/

Thank You

Rare Words

- Performs well with single training examples

Singleton training example



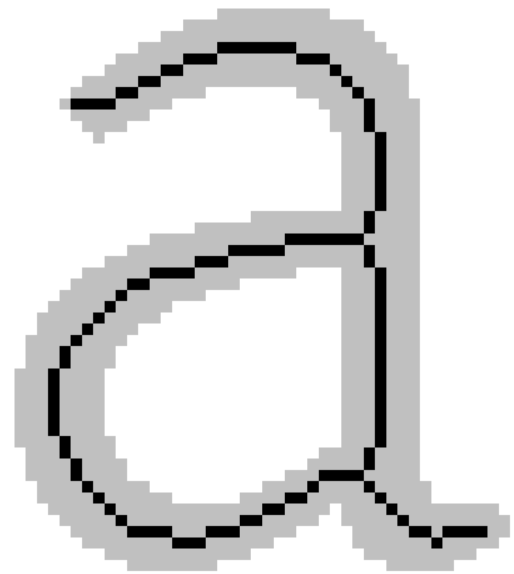
All three instances in target set rank at top of retrieval list

GW20: 25.4% of queries are singletons ➔ 60.2% precision at full recall

Parzival: 31.8% of queries are singleton ➔ 69.5% precision at full recall

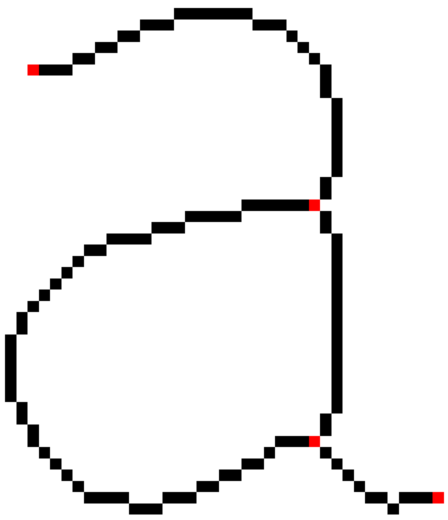
Building PSM from Image

1. Find skeleton



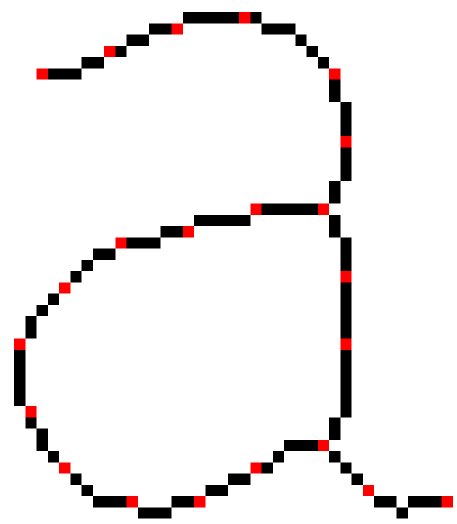
Building PSM from Image

1. Find skeleton
2. Select endpoints & junctions



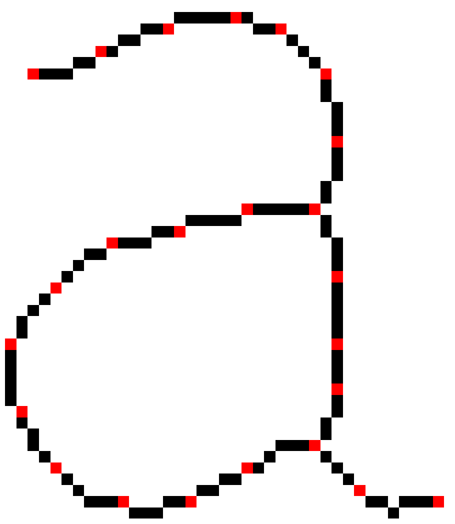
Building PSM from Image

1. Find skeleton
2. Select endpoints & junctions
3. Add points chosen ***2r*** from included points



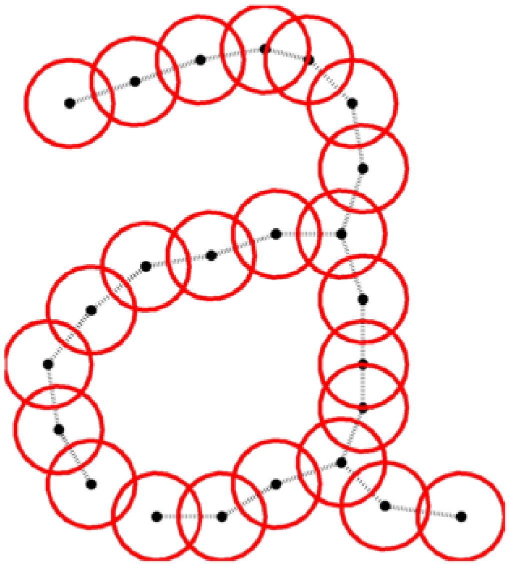
Building PSM from Image

1. Find skeleton
2. Select endpoints & junctions
3. Add points chosen ***2r*** from included points
4. Additional points to fill remaining gaps



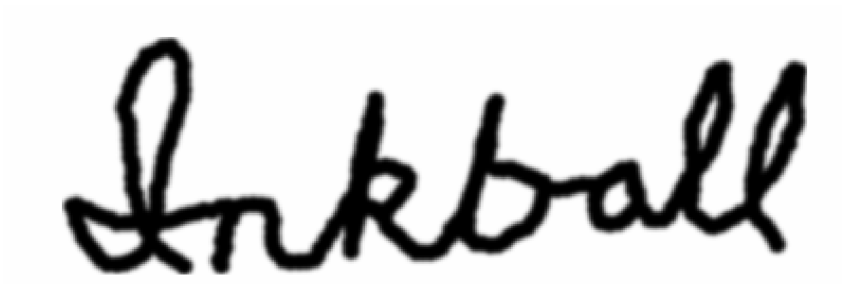
Building PSM from Image

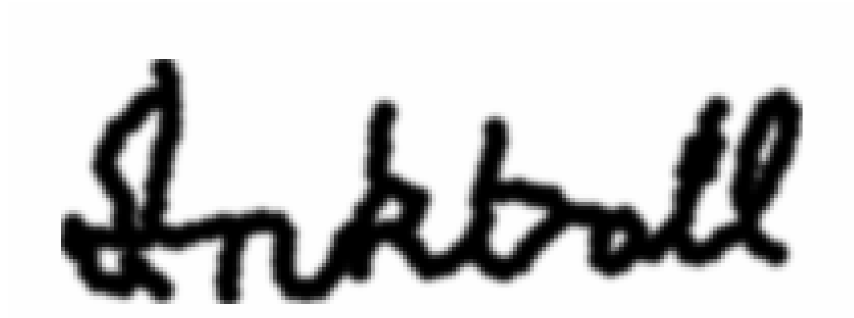
1. Find skeleton
2. Select endpoints & junctions
3. Add points chosen ***2r*** from included points
4. Additional points to fill remaining gaps
5. Form tree by greedily connecting closest pairs



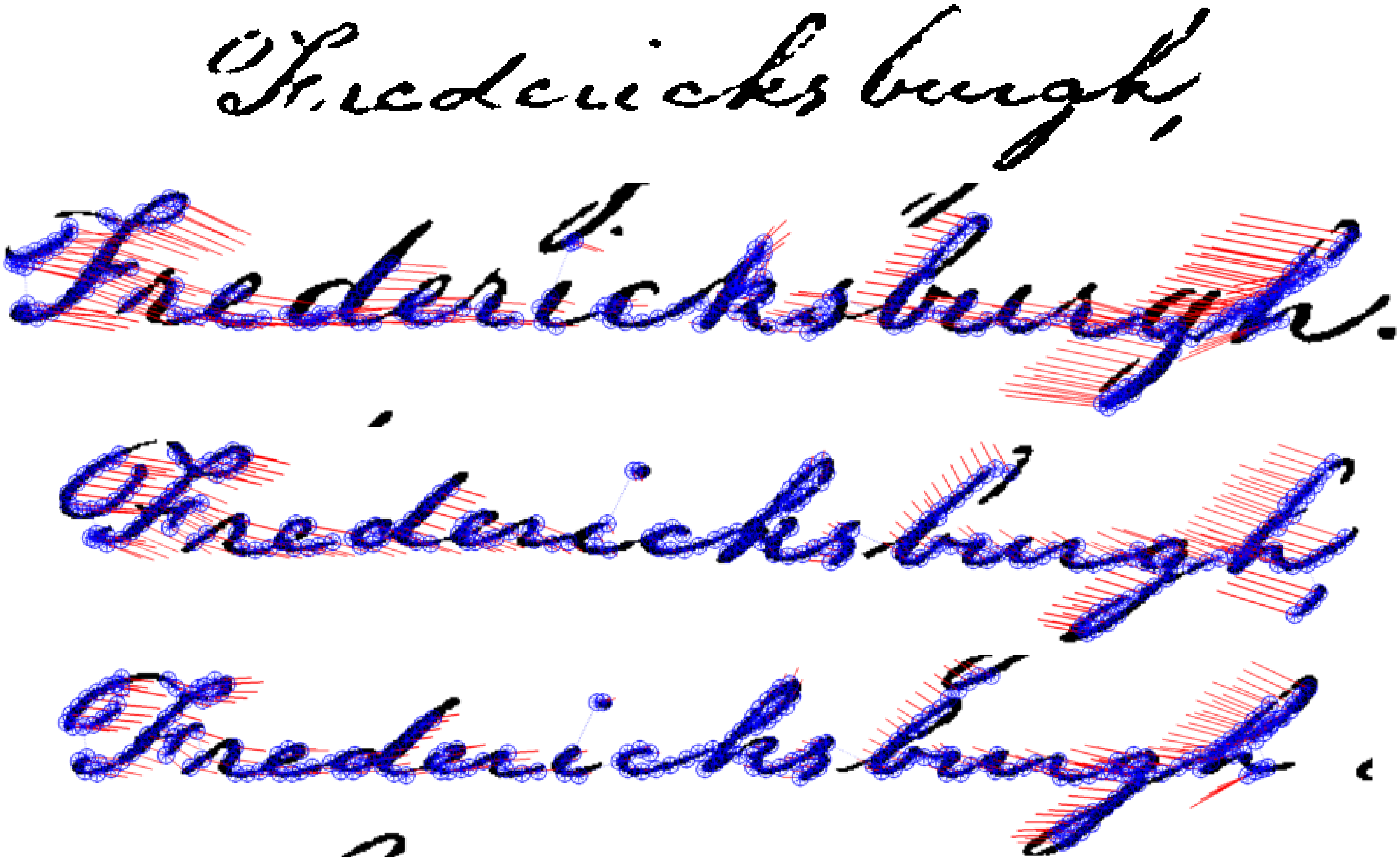
Online vs. Offline Models

- Online query allows model structure to follow actual stroke
- Offline query must use *ad hoc* model structure

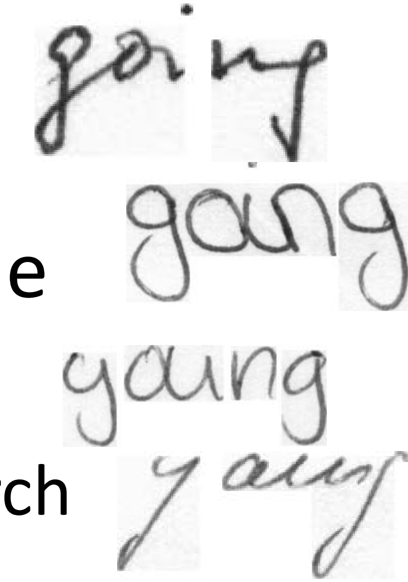




Some Matches



Caveat Lector



- Some dependence on handwriting style
  - – Intrinsic letter forms can vary
  - – Cross-style spotting requires more research
- Limited invariance to scale & rotation
  - – Match model scale to text in document
  - – Correct skew/rotation prior to spotting
- Speed not yet real-time for large collections
  - – Roughly 2-3 Mpixel/second for most words